CHARACTERIZING THE HUMAN INTESTINAL MICROBIOME

CHARACTERIZING THE HUMAN INTESTINAL MICROBIOTA IN HEALTHY INDIVIDUALS AND PATIENTS WITH ULCERATIVE COLITIS USING CULTURE-DEPENDENT AND -INDEPENDENT APPROACHES

By

SHAHROKH SHEKARRIZ, M.SC., B.SC.

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University

Doctor of Philosophy (2022)

Hamilton, Ontario (Department of Biochemistry & Biomedical Sciences)

TITLE: CHARACTERIZING THE HUMAN INTESTINAL MICROBIOTA IN HEALTHY INDIVIDUALS AND PATIENTS WITH ULCERATIVE COLITIS USING CULTURE-DEPENDENT AND -INDEPENDENT APPROACHES

AUTHOR: Shahrokh Shekarriz (McMaster University)

SUPERVISOR: Dr. Michael G Surette

NUMBER OF PAGES: xix, 194

Lay Abstract

Many bacteria reside in the human gut, and they are essential in our health and in disease. It is evident that these bacteria are associated with inflammatory bowel disease, but we do not yet know how and what bacteria are involved in this disease. In this work, I describe a method to study these bacteria from stool that relies on growing them and investigating their DNA. I showed that our approach helped us recover a greater diversity of these bacteria and their genetic content in healthy individuals and patients with inflammatory bowel disease compared to methods that use only DNA based approaches. Using this method, we could better understand why some patients responded to a treatment consisting of transferring stool content from healthy donor to patient. I also investigated a group of viruses that infect bacteria and implemented a new computational method based on DNA sequencing to test whether these viruses transfer to the patient after receiving the fecal therapy. We also found that antibiotic treatment before fecal therapy in patients with inflammatory bowel disease does not improve the patient's recovery.

Abstract

The collection of microbes that inhabits the human gastrointestinal tract is known as intestinal microbiota, and an enormous body of work has shown that their activities contribute to health and disease. Ulcerative colitis (UC), which is a type of inflammatory bowel disease, is considered to arise due to a disruption in the balance between the immune system and microbiota. However, there is little consensus on the mechanism of action and microbes involved in the disease manifestation. In this work, I applied culture-enriched metagenomics (CEMG) to characterize the dynamics of gut microbiota in healthy individuals and UC patients. I showed that CEMG provides a higher resolution to study these microbial communities, and we used this approach to understand microbial colonization after fecal microbiota transplantation (FMT) therapy in UC patient. I showed that sequencing approaches alone did not reveal consistent engraftment across FMT responders. Using CEMG and a collection of bacterial whole-genome sequences, I showed patient-specific microbial strain transfer and a signature of commonly engrafted genes only in patients who responded to FMT. In this work, I also investigated the dynamics of a highly abundant bacteriophage, crAssphage, in an FMT donor and implemented a new method to detect bacteriophage engraftment post-FMT using SNP analysis. Finally, it has been suggested that antibiotic treatment before FMT may increase the efficacy of FMT. However, in this work, I show that while antibiotics alter the microbiome, there was no difference in the composition of the microbiome of antibiotic vs placebo group post-FMT. This is consistent with the randomized controlled trial results that shows pretreatment with antibiotics does not improve FMT outcome. Together, this work demonstrate the importance of in-depth microbiome analysis applied to culture-dependent and -independent sequencing to characterize microbial changes post-FMT.

Acknowledgements

I would like to thank my supervisor, Dr. Michael Surette, for giving me the opportunity to work in his lab and inspiring me as a scientist. He has created an environment that allowed me to grow my scientific curiosity and encouraged me to challenge thoughts and ask questions. In a world full of distractions, his passion for science has been a source of encouragement and gave me the confidence to learn new ways, get excited by new data, and even enjoy failing in science. All the work presented in this document is the result of my discussions with Mike, and I can not imagine completing my doctoral work without his support. As an international student, I had many challenges during my doctoral work, but his mentorship, support, and interest in my career development made this journey much easier.

I would also like to thank my committee members, Dr. Henrik Poinar and Dr. Andrew McArthur, for guiding me during my doctoral work. I really appreciate their excitement about my projects, challenging questions, and their brilliant suggestions. I also thank Dr. Andrew McArthur for providing computational resources whenever we needed extra power. I would like to thank Dr. Paul Moayyedi for providing us with invaluable data, clinical insights, and his interest in my career development.

Thank you to all the current and past members of the Surettelab for their support and help. Thank you for all you have taught me and shared with me. My graduate school experience has been great because of your company, thoughtful scientific discussions, excitement about my projects, and even the jokes you made about my skills to press "enter". To Laura Rossi, thank you for listening to all my daily rants, helping me sort out issues, and being a source of wisdom when I needed the most. Thank you, Michelle Shah, for all your work for UC-FMT studies and for supporting me throughout my Ph.D. work. To Jake Szamosi, thank you for all your arguments with me about the right figure colours, teaching me stats, providing feedback on my projects, and making the Farncombe office so much more fun. Thank you, Dr. Fiona Whelan, for all the brainstorming skype calls, and voice memos during the pandemic. We had a short overlap in the lab, but you were a source of inspiration and encouragement for me. Thank you, Saad Syed, for your company, scientific brainstorming and pandemic lockdown walks. You made the grad school experience much more fun for me.

Thank you to my wonderful family, who were away but supported me as much as they could during this journey. Thank you, Mom, Khalejon, Mamany, Shiva, and Shaghayegh, for your endless love. Your presence has always comforted me and encouraged me to pursue my dreams. Thank you to my cousin, Sheida, for all the heart-to-heart chats and tremendous support during the past few years. Last but certainly not least, thank you to my cousin, Dr. Mehrdad Hajibabaei, my big brother. You have shown me the way, supported me by all means, and have been a great role model for me since I remember. Thank you for your gift; the scientific outlook book by Bertrand Russell profoundly changed my perspective as a teenager and encouraged me to become a scientist.

Contents

List of Figures

List of Tables

Abbreviations

AMR antimicrobial resistance

ASV amplicon sequence variant

C. difficile *Clostridioides difficile*

CDI *Clostridioides difficile* infection

CEMG culture-enriched metagenomics

COG cluster of orthologous group

CD Crohn's disease

DMG direct metagenomics

E. coli *Escherichia coli*

FMT fecal microbiota transplantation

GF germ-free

GI gastrointestinal

HAIs heathcare associated infections

HMO human milk oligosaccharide

HMP Human Microbiome Project

HTS high-throughput sequencing

IBD inflammatory bowel disease

IBS irritable bowel syndrome

IMM interpolated Markov model

LCA lowest common ancestor

MAG metagenome assembled genome

MetaHIT Metagenomics of the Human Intestinal Tract

MLST multilocus sequence typing

ORF open reading frame

OTU operational taxonomic unit

PCR polymerase chain reaction

PSA Polysaccharide A

rCDI recurrent- *Clostridioides difficile* infection

RCT randomized controlled trial

SCFA short chain fatty acids

SNP single-nucleotide polymorphism

SPF specific-pathogen-free

SRA sequence read archive

UC ulcerative colitis

WGS whole-genome sequencing

zOTU zero-radius OTU

Declaration of Authorship

I, Shahrokh Shekarriz, declare that this thesis titled, "CHARACTERIZING THE HUMAN INTESTINAL MICROBIOTA IN HEALTHY INDIVIDUALS AND PATIENTS WITH ULCERATIVE COLITIS USING CULTURE-DEPENDENT AND -INDEPENDENT APPROACHES" and the work presented in it are my own.

Chapter 1

Introduction

1.1 The human microbiome

The community of microbes (bacteria, viruses, archaea, and fungi) that inhabit the human body is known as microbiota, and their "theatre of activity" modulated by genotype —genetic makeup — is referred to as the microbiome (Whipps et al. 1988; Berg et al. 2020). The number of microbial cells is as abundant as the somatic cell in humans (Sender et al. 2016). These microbes collectively contain more genetic content than the human genome. Estimates vary and are often exaggerated, but hundreds of microbial species, with each genome containing at least 1000 genes, live on and inside the human body (Locey and Lennon 2016), while there are 19-22 thousand host genes (Willyard 2018).

The human microbiome has been studied since the seventieth century (1670) with Antonie van Leeuwenhoek's work on discovering microorganisms, which he called "animalcules". In 1884, Robert Koch elucidated the concept of pathogenicity and defined microbial infection as the cause of human diseases. Although this definition was an essential milestone in microbiology, it has shaped the role of microorganisms as harmful

agents. Today, most members of the human microbiota are considered commensal- generally beneficial, if not essential, organisms that do not harm their host (Hugon et al. 2015).

The commensal microbes have co-evolved with their animal hosts for millions of years and have become highly adapted to the specific host niches (Dominguez-Bello et al. 2019). For example, *Bifidobacterium* acquired during birth are essential for neonate development. These bacteria use human milk oligosaccharides (HMOs) that are indigestible for infants as energy sources, and they facilitate babies' immune, metabolic and nervous system development (Zivkovic et al. 2011; Hamilton et al. 2017; Berger et al. 2020). Another example of mutualistic co-adaptation between commensal microbes and the host is colonization resistance- the mechanisms microbiota uses to protect against invasion of exogenous pathogens in their host (Levine and D'Antonio 1999; Hibbing et al. 2010; Lawley and Walker 2013).

The two initiatives, Human Microbiome Project (HMP) and Metagenomics of the Human Intestinal Tract (MetaHIT), contributed significantly to the understanding of the microbiome associated with the healthy human (Turnbaugh et al. 2007; Ehrlich, Consortium, et al. 2011). These programs revealed a tremendous microbial heterogeneity between healthy individuals and between body sites. The cross-sectional samples from healthy individuals were primarily clustered based on body sites, suggesting that the microbial communities residing in a body site (e.g. oral cavity, vaginal, lung, etc.) from different individuals were more similar than communities present in multiple body sites of the same individual. Longitudinal samples from the same individuals were more similar than samples from healthy individuals, highlighting heterogeneity among individuals and stability of the healthy microbiome over time. HMP and MetaHIT examined only samples from developed ("westernized") countries. However, other studies from indigenous communities showed that the non-western microbiomes consists of more species

and an increase in the relative abundance of *Firmicutes* and *Proteobacteria* (De Filippo et al. 2010; Yatsunenko et al. 2012; Schnorr et al. 2014; Clemente et al. 2015).

1.2 The human gut microbiome

The largest microbial community in the human body inhabits the gastrointestinal (GI) tract, from the mouth to the anus, and they play a fundamental role in health. A growing body of evidence suggest that the human gut microbiome is shaped by internal and external factors such as host genetics (Benson et al. 2010; New et al. 2022), geography (Deschasaux et al. 2018), diet (David et al. 2014), and disease state (Greenblum et al. 2012). Further, age affects the composition of the intestinal microbiota (O'Toole and Jeffery 2015), and aspects including the maternal microbiome (Mueller et al. 2015), mode of delivery (Dominguez-Bello et al. 2010), and antibiotics (Bokulich et al. 2016) shape the early life microbiome in humans.

The gut microbiome modulates many critical functions, including fermentation of indigestible dietary compounds, such as fibres, into short chain fatty acids (SCFA). Butyrate that is a SCFA that enhances the intestinal barrier and has anti-inflammatory properties (Morrison and Preston 2016; Peng et al. 2007; Maslowski et al. 2009). It was shown that the abundance of *Lachnospiraceae*, a butyrate-producing bacterial family is reduced in patients with inflammatory bowel disease (IBD), suggesting the importance of these bacteria for modulating digestion and producing metabolites (Frank et al. 2007; Morgan et al. 2012). The microbiome is also actively involved in protecting against pathogens (Kamada et al. 2013; McDonald et al. 2020), and stimulating the immune system (Wu and Wu 2012; Maynard et al. 2012). Germ-free (GF) and specific-pathogen-free (SPF) mice studies have helped us to appreciate the delicate balance and interactions between intestinal microbiota and the immune system (Hooper et al. 2012). Microbial residence in the intestine shape systemic immunity by mediating regulatory T

cells that maintain immune homeostasis and inflammation. For example, Polysaccharide A (PSA) produced by *Bacteroides fragilis* directly affect regulatory T cell activity via TLR2 signalling of dendritic cells (Round and Mazmanian 2010; Shen et al. 2012; Smith et al. 2013).

1.3 The human gut microbiome in disease

Dysbiosis — loosely defined as disease-related disruption of microbiota — of the intestinal microbiota has been implicated in both GI-related and non-related diseases. Although the term "dysbiosis" is often used to describe a deregulated microbial community without considering that the healthy microbiota is highly heterogeneous, it is clear that a shift in the microbial community of the intestine is associated with disease and disorder in human (Shanahan et al. 2021). IBD, irritable bowel syndrome (IBS), and colorectal cancer all have been associated with the altered gut microbiota (Zhang et al. 2022; Ford et al. 2018; Pleguezuelos-Manzano et al. 2020). Further, the balance and composition of the intestinal microbes have been shown to affect depression, Parkinson's disease, and autism disorder through a more complex system called the gut-brain axis (Bastiaanssen et al. 2019; Sampson et al. 2016; Sharon et al. 2019). These microbial disruptions could be manifested in the relative abundance of a diverse group of bacterial phyla, specific strains, and functional changes. Recognizing the cause-and-effect relationship between the intestinal microbiota and other factors in the context of a disease is also essential. Microbiome changes could be a consequence or a cause of disease. For example, it was shown that the chemicals used to induce inflammation could cause altered gut microbiota in mice (Lupp et al. 2007) and this change in microbiome may be a consequence of disease induction. On the other hand in vivo transfer of gut microbiota from patients with IBS could recapitulate disease phenotypes in naive mice (De Palma et al. 2017) supports causation. Sample size is another critical factor in finding whether

there is a direct association between a disease and microbiome and adequately comparing it to other confounding factors. For example, recently, it was implicated that dietary preferences caused by autism modulate the microbial changes in these patients (Yap et al. 2021).

1.4 Inflammatory bowel disease

IBD, defined as chronic inflammation of the GI tract, includes two similar but distinct conditions: ulcerative colitis (UC), and Crohn's disease (CD). The etiology of IBD is unknown, but it is known to be caused by a complex interplay between host genetics, environment and the immune system. The incidence of IBD is rapidly rising in developed countries, especially in Canada, and particularly in children (Ng et al. 2017; Benchimol et al. 2009). Over 200 genes have been link to IBD (such as *NOD2*, *ATG16L1*) involved in epithelial barrier integrity, autophagy, and oxidative stress (Imielinski et al. 2009; Hampe et al. 2007; Hugot et al. 2001). However, only 20% of IBD cases are explained by genetics (Peters et al. 2017), and the recent increase in IBD incidence can not reflect genetics alone and highlights the importance of environmental factors.

Diet (Levine et al. 2018; Rangan et al. 2019; Liu et al. 2021), smoking (Mahid et al. 2006), infection in infancy (Bernstein et al. 2019), and the gut microbiome are environmental factors associated with IBD. These factors are not all equally important and seem to depend on the study population and sample size. For example, it was shown that smoking increases the chance of CD but reduces the risk of UC (Calkins 1989). The gut microbiota is the most potent environmental factor related to IBD reproduced in meta-analysis, microbiome analysis from population studies, and mice models (Llewellyn et al. 2018; Walters et al. 2014; Abbas-Egbariya et al. 2022; Franzosa et al. 2019; Lee et al. 2021).

1.4.1 Crohn's disease

Chronic inflammation in CD is patchy, asymmetrical, transmural and can affect all segments of the GI tract. CD is associated with dis-regulated barrier function due to increased intestinal permeability (Torres et al. 2017). The impaired intestinal barrier in CD results in leaky tight-junction and loose regulation of transepithelial transport that allows pathogenic bacteria to induce immune responses that can lead to intestinal inflammation (Libertucci et al. 2018). It was inferred that the reduction of butyrate-producing bacteria (i.e. *Clostridia*) in CD leads to increased O_2 in the gut lumen by intestinal epithelial cells. As a result of this change in oxygen level, facultative anaerobe (e.g. *Escherichia coli*) expansion and the loss of obligate anaerobes accelerates (Byndloss et al. 2017; Rivera-Chávez et al. 2016; Mottawea et al. 2016). The overgrowth of *E. coli* strains, that adhere to ileal tissue using the FimH adhesin, has been shown in CD patients and hypothesized to be one of the causes of CD (Lapaquette et al. 2012; Martinez-Medina and Garcia-Gil 2014).

1.4.2 Ulcerative colitis

UC is a chronic disorder characterized by inflammation and ulceration of the colonic mucosa. Canada has one of the highest incidents of UC worldwide, with a peak incident in early adulthood (Molodecky et al. 2012; Ng et al. 2017). The primary symptoms of UC are bloody diarrhea, abdominal cramps, fatigue, increased risk of colon cancer, and increased depression, which significantly impact the quality of life (Collins et al. 2012). The cause of UC is unknown, but it is generally considered that the disease arises from an immune response to altered intestinal microbiota in genetically susceptible individuals (Talley et al. 2011). Antimicrobial peptide secretion, antigen presentation, and intestinal barrier are reduced in UC patients and contribute to increased inflammation (Ho et al. 2013). Although UC-related microbial changes are less known than CD, it was implicated that UC patients' ability to produce SCFA is diminished, and microbial diversity in these

patients is reduced compared to healthy controls (Michail et al. 2012; James et al. 2015).

Current therapies for UC are primarily focused on suppressing the immune response with anti-tumour necrosis alpha monoclonal antibodies (anti-TNFα), 5-aminosalicylic acid (5-ASA), and corticosteroid therapy (Talley et al. 2011) rather than reducing factors that stimulate immune response(Danese 2012). As a result, these immune suppressive treatments are associated with increased risk of infection (Tinsley et al. 2013; Kirchgesner et al. 2018) and colon cancer (Ekbom et al. 1990; Eaden et al. 2001). Prescription drugs accounts for 42 % of total direct costs for IBD patients in Canada, and costs to treat IBD continue to rise due to increased use of existing biologic therapies and the introduction of several new biologic therapies in recent years. For example, in Manitoba, the mean healthcare utilization and medication costs for persons with IBD in the year before beginning anti-TNF treatment was $10,206 and increased to $44,786 in the first year of therapy (Crohn′s and Colitis Canada 2018).

If the altered colonic microbiome is the trigger of immune responses, then alternative treatments are required to restore microbiota-intestinal immune homeostasis. Antibiotic therapy and fecal microbiota transplantation (FMT) are microbiome targeting therapies that have been trialled for UC patients. A systematic review focused on the efficacy of antibiotics versus placebo showed that antibiotic treatment had a modest effect on patients with UC (Khan et al. 2011). However, they could not make any recommendations because different antibiotics were used in every trial. These therapies and their efficacy in UC are part of the main objectives of this thesis, and I will discuss these therapies in Chapters 3 and 4.

1.5 Fecal microbiota transplantation

1.5.1 History of FMT

FMT — administration of a fecal suspension from a healthy donor to a patient — is an ancient therapy that goes back to 1,700 years ago. In the 4th century, Ge Hong in China used FMT to treat food poising and diarrhea. Li Shizen, in the 16th century, referred to FMT as "golden syrup" to treat patients with abdominal pain, diarrhea and even fever (Zhang et al. 2012; De Groot et al. 2017). It is unclear how effective these treatments were and how they originally started long before discovering microbes. It is possible that the idea behind this treatment was first created by observing animal species that naturally practice coprophagia, potentially enabling them to have a more diverse diet. Numerous other cases of FMT historically reported in different diseases, particularly in veterinary medicine (Mullen et al. 2018), but the first modern study conducted for four patients with pseudomembranous colitis, likely caused by *Clostridioides difficile* infection (CDI), resulted in complete recovery for all participants (GS, AJ, et al. 1958).

1.5.2 FMT in CDI

CDI is one of the leading cause of heathcare associated infections (HAIs) in the world (Khanna et al. 2012). Ubiquitous spores of *Clostridioides difficile* (*C. difficile*) can stay infectious for a long time and can transfer to GI tract of both animals and humans (Paredes-Sabja et al. 2014). Antibiotic agents, metronidazole and vancomycin, are the standard therapy for CDI (Shen and Surawicz 2008; Bagdasarian et al. 2015). The risk of complication associated with CDI increases by antibiotic use and age possibly due to microbial changes that may result in loss of colonization resistance. An episode of CDI occurring within two months of the initial infection either by the same or different strain is known as recurrent- *Clostridioides difficile* infection (rCDI). It is estimated that 15-30% of patients who initially respond to antimicrobial therapy will develop rCDI (Song and Kim 2019). Currently, FMT is the standard treatment for rCDI with $\geq$ 90%

remission rate. (Kassam et al. 2012; Van Nood et al. 2013; Quraishi et al. 2017). It was implicated that the microbial community shift modulated by antimicrobials depletes bile acid production and promotes *C. difficile* growth in the large intestine of patients with rCDI (Theriot et al. 2016). In this context, FMT can bring back a more diverse microbial communities to the intestine and potentially increase colonization resistance.

1.5.3 FMT in CD

FMT has been used for the treatment of CD since 1989 (Borody et al. 1989), but its efficacy remained controversial because the reported studies contained small participants and lacked proper controls (Cui et al. 2015; Suskind et al. 2015; Vaughn et al. 2016). More recently, the first pilot randomized controlled trial (RCT) aimed to investigate the efficacy of FMT for CD (FMT:n=8 vs. Placebo:n=9) showed that the clinical remission at 10 weeks was 87.5% in the FMT group compared to 44.4% in the sham transplantation group. Further, Yang et al. 2020 conducted a RCT and showed that there was no significant difference in delivering FMT via gastroscopy and colonoscopy in the small intestine and colon, respectively. Systematic reviews suggest that FMT is a safe and potentially effective treatment, but further randomized clinical trials are needed to evaluate their efficacy in CD comprehensively (Cheng et al. 2021; Fehily et al. 2021).

1.5.4 FMT in UC

Similar to CD, the first case of FMT in UC was reported in 1989 (Bennet and Brinkman 1989), but FMT has shown to be more successful in UC than CD. Six adult and one pediatric RCTs have been conducted so far to evaluate the efficacy of FMT in UC. Rossen et al. 2015 concluded that there was no significant difference between the FMT and placebo group in clinical remission at the end of their study. They included UC patients with mild disease activity, which resulted in an increased remission rate in the placebo group. The first successful RCT was conducted in Canada, showing that

24% of patients who received FMT went into remission versus 5% in the placebo group (Moayyedi et al. 2015). Since then, these results have been reproduced in multiple other trials, and they all confirmed the safety and efficacy of FMT treatment in UC (Paramsothy et al. 2017; Costello et al. 2019; Smith et al. 2022; Haifer et al. 2022b). FMT was delivered to the patient's colon via enema in all of these RCTs except two studies (Smith et al. 2022; Haifer et al. 2022b) that used oral FMT capsules. A study randomizing to colonoscopy versus lyophilized pills was too small to measure a deference between delivery methods (Crothers et al. 2018).

1.5.5 FMT donors

Selecting an appropriate donor for the FMT studies has been controversial, and still not clear whether the microbial composition of the donor determines the FMT success. For example, Van Nood et al. 2013 studied FMT donors for CDI patients suggested that there was no apparent difference between donors; however, FMT outcomes from UC (Moayyedi et al. 2015) and obesity (Wilson et al. 2021) patients implicated that the choice of the donor is important. It was shown that the donor's species richness (Vermeire et al. 2016), metabolite fitness (Watson et al. 2021), and stability (Haifer et al. 2022a), are likely important factors that indicate a successful donor. At the same time, more extensive studies that merged various datasets of different diseases suggested that the recipient's factors outweigh the donor's microbiome composition (Schmidt et al. 2022). Further, it was recommended that matching a recipient to a suitable donor should be the priority in selecting successful donors (He et al. 2022). Given the variation between these diseases, the importance of donor may be disease-dependent related to the mechanism of FMT.

1.5.6 Mechanism of FMT

Microbial engraftment — donor's microbiota that transfer and colonize in the FMT recipient — is considered as the main mechanism of FMT and focuses on restoring the bacteria in the gastrointestinal tract (Youngster et al. 2014) that may change host metabolism (Floch 2015), host immunity (Furusawa et al. 2013; Round and Mazmanian 2010), and restrain pathogens (Britton and Young 2014). This mechanism of action is likely disease dependent. For example, CDI is an acute infection disease while IBD is a chronic inflammatory disease. The goal of FMT in rCDI is to restore the community balance but in IBD it needs to fix the metabolic dysfunction. Previously, it was shown that the donor-specific bacteria might establish alongside the host microbiota, and they can be detected after FMT (Angelberger et al. 2013; Fuentes et al. 2014), but it is difficult to determine if these newly observed bacteria are transferred from the donor or present in the patient prior to treatment at low levels, meaning that the new detected bacteria had been below detection level in recipient before FMT and they became more abundant after FMT.

In addition to the donor, other factors can potentially influence FMT outcome. These factors include mode of delivery, anaerobic considerations for FMT preparation, duration of FMT treatments, and antibiotic pretreatment. The FMT treatment may proceed by a course of antibiotics that presumably alter the intestinal-microbiota (Dethlefsen and Relman 2011) and may facilitate the implantation of donor-specific bacteria. During inflammation, immune cells Increase their uptake of oxygen, reducing oxygen levels at the epithelial layer (Campbell et al. 2014). As a result of these oxygen changes, epithelial cell absorption and barrier functions are disrupted (Rigottier Gois 2013). Considering the important role of obligate anaerobes in gut homeostasis (Peterson and Artis 2014), particularly during inflammation, preserving these microbes during FMT preparations should be prioritized.

1.6 Studying the gut microbiome

1.6.1 16S rRNA gene amplicon sequencing

One of the fundamental research aims of studying gut microbiota is to uncover the composition and the abundance of microbiota. 16S ribosomal RNA (rRNA) gene amplicon sequencing provides a relatively cost-efficient approach to estimating the bacteria's abundance in a sample. 16S rRNA gene has been traditionally used to determine the phylogeny of prokaryotes (Fox et al. 1977). The 16S rRNA gene with a total length of ∼1500bp is a highly conserved gene containing nine variable regions (V1 - V9), which makes it suitable for primer binding and capturing diverse bacteria (as well as some archaea depending on the variable region) (Woese et al. 1990). These regions within the 16s rRNA gene are typically referred to as hypervariable regions, and universal primers have been used to amplify these regions, such as the variable 3, 4 and 5 regions (Caporaso et al. 2010b). High-throughput sequencing (HTS) technologies, most notable of which being the second-generation platforms such as Illumina, have provided the ability to sequence regions up to 600 bp on a large scale. More recently, the third generation platforms (e.g. PacBIOs Sequel and Oxford Nanopore MinION) allowed sequencing of the entire 16S gene, but they lack standardization and MinION include a relatively high error rate (Rhoads and Au 2015; Bowden et al. 2019).

The 16S rRNA gene sequencing workflow includes clustering sequences into operational taxonomic units (OTUs) at 97% sequence similarity. Alternatively, a 100% sequence identity threshold could be used, by implementing denoising methods, to identify amplicon sequence variants (ASVs) or zero-radius OTUs (zOTUs) (Callahan et al. 2016; Edgar 2018). Next, these clustered sequences are used for taxonomic classification using different programs (e.g. Mothur (Schloss et al. 2009), Qiime (Caporaso et al. 2010a), Qiime2 (Bolyen et al. 2019), DADA2 (Callahan et al. 2016), etc.) and databases (e.g. GreenGenes (GG) (DeSantis et al. 2006), the Ribosomal Database Project (RDP) (Cole

et al. 2014), Silva (Quast et al. 2012), etc.).

In order to reproduce microbiome findings, the sequencing pipelines need to be standard, meaning that the difference between datasets represents biological differences and not technical variations. Szamosi et al. 2020 found that 16S rRNA gene amplicon analysis variation in the extraction and sequencing protocols are less sensitive than data processing pipelines when they compared matched samples processed in multiple laboratories, suggesting the importance of bioinformatics workflows. The factors that contribute most to variations in 16S rRNA gene analysis include the choice of primer (variable region), reference databases, and to a less extent, clustering approaches (Abellan-Schneyder et al. 2021). The other caveat in 16S rRNA gene sequencing is the low taxonomic resolution. Although debatable and depending on the taxonomic group, the identified OTUs and ASV represent bacterial genera occasionally accurate to the species-level (Johnson et al. 2019) which seems insufficient to study intestinal microbiota given the observed species and strain variations (Truong et al. 2017; Park et al. 2022).

1.6.2 Read-based metagenomics

The 16S rRNA gene sequencing led to the discovery of novel microbial diversity, but the lack of culture representatives for many microbial groups, such as Archea, demanded a new approach to investigate these microbes. Stein et al. 1996 reported the first attempt to solve this problem by random shotgun sequencing of the archaeal clones extracted from picoplankton assemblage collected in the Pacific Ocean. However, the term *metagenome* was used two years later to refer to "the collective genomes of soil microflora" (Handelsman et al. 1998). Since then, "metagenomics" have been used to describe various data structures. For example, 16S rRNA gene amplicon sequencing sometimes is referred to as metagenomics inaccurately- maybe because this approach could identify microbes beyond one genome (Arboleya et al. 2012; Brooks et al. 2015). Despite issues with metagenomic terminology, shotgun (untargeted) metagenomics is trying to uncover

both "what is there" regarding the functional potential of microbial community and "who is there" regarding microbiota composition.

In the last decade, the reduced cost and improvement in DNA sequencing have allowed large-scale metagenomics to study human microbiota (Temperton and Giovannoni 2012). The higher taxonomic and functional resolution in metagenomic sequencing has significantly improved our understanding of the human microbiota. In this approach, the total DNA will be extracted from a sample (e.g. fecal, biopsy, swab, etc.) and a sequencing library will be prepared depending on the sequencing technology platform. Currently, the most common sequencing platform for metagenomic sequencing is Illumina (HiSeq, NextSeq, and NovaSeq), which generate 150-250bp sequence reads. PacBIO and Nanopore can sequence a longer DNA fragment but is less frequently used due to the higher cost (Sevim et al. 2019; Mahmoud et al. 2019).

Read-based metagenomics aims to profile a microbial community's taxonomy and functional capacity without necessarily gaining knowledge of the microbial members that contribute to the functions or genes that are present but not annotated in the publicly available databases. This approach compares the reads that passed the quality control to external sequence databases. There are three main approaches that compare query sequence to databases for taxon and/or functional assignment (supervised learning): similarity search (use homology or alignment-based methods based on lowest common ancestor (LCA) ; e.g., BLAST (Altschul et al. 1997) and MEGAN (Huson et al. 2011)), composition methods (use k-mer counts or frequencies; e.g., KRAKEN (Wood and Salzberg 2014), RDP (Wang et al. 2007)); and phylogenetic approach (use evolutionary models coupled with homology-based or interpolated Markov models; e.g., (Brady and Salzberg 2009)). The homology-based method searches each query sequence against large databases that takes a long time. The phylogenetic approach for taxonomic classification employs evolutionary models utilizing maximum likelihood, neighbor-joining, or

Bayesian methods to calculate the suitable place of a query sequence on a phylogenetic tree (Bazinet and Cummings 2012). These tools use simple observation to find where an inserted branch is divergent from a node representing a species or higher rank. It requires enormous computational power as it contains multiple alignments, fixed topology (e.g., NCBI taxonomy), and the insertion of a query sequence into the reference alignment. The compositional methods, including the Naive Bayesian classifiers, interpolated Markov models (IMMs) and kmer/k-nearest-neighbor algorithms (Ames et al. 2013) are much faster than alignment or phylogenetic-based approaches. Still, they require a large computational memory because a pre-computed database needs to be pre-loaded into the memory.

Marker-based algorithms are another read-based approach that incorporates a set of representative genes (markers) instead of a more extensive database of all known sequences to profile microbial composition. These assembly-free methods have been used to analyze large human associated metagenomic datasets from MetaHIT and HMP consortiums via mOTU (Sunagawa et al. 2013) and MetaPhlAn (Segata et al. 2012; Beghini et al. 2021), respectively (Voigt et al. 2015; Nielsen et al. 2014; Lee et al. 2022). For example, it was shown that the clade-specific markers in the CHOCOPhlAn database, used in MetaPhlan, provide an accurate estimate of microbial composition and, most importantly, offer a faster run time (Meyer et al. 2021). The main caveat is to profile previously unknown microbes, particularly gene families and functions. For example, HUMAnN package that is being used to profile functional pathways and gene families usually returns 40% unmapped reads (Franzosa et al. 2018). Although the list of reference genomes is exponentially expanding and markers are becoming more accurate in detecting species and pathways, these databases are often not well annotated or complete.

1.6.3 Assembly-based metagenomics

In 1995, the two bacterial genomes *Haemophilus influenzae* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (Fraser et al. 1995) were completely sequenced. Since then, DNA sequencing technologies have revolutionized systems biology and biomedical research. Recent advancements significantly reduced the cost of sequencing and resulted in a dramatic growth of genomic data from all organisms, particularly human microbiota (Muir et al. 2016). Despite these advancements, current technologies can only sequence small genomic fragments, ranging from 150bp (such as Illumina) to approximately >10–20kb (such as PacBIO). A typical bacterial genome is 5 million bp (Land et al. 2015); thus, reconstructing the whole genome requires a sophisticated computational algorithm to assemble the short sequencing reads together. By contrast, human intestinal microbiota contains thousands of these bacterial genomes making the gut metagenome assembly a daunting task.

The two main genome assembly approaches include reference-based and *de novo* assembly (reference-independent). Due to the diversity of the healthy gut microbiome (Lozupone et al. 2012) and the incomplete nature of microbial reference databases (Loeffler et al. 2020), it is essential to reconstruct the metagenome structure of the new microbes in an unbiased reference-free approach. Although there have been some attempts to use reference-guided methods (Dutilh et al. 2009; Tsai et al. 2010; Lischer and Shimizu 2017), predominantly *de novo* assemblers were used to assemble microbial genomes and metagenomics (Quince et al. 2017).

The *de novo* assembly approach can be classified into three basic categories: OLC graph, string graph, and de Bruijn graph. The OLC algorithms (such as Celera (Myers et al. 2000), AMOS (Treangen et al. 2011), and PCAP (Huang et al. 2003)) work based on three main principles: finding overlaps across the reads, constructing a layout graph from the overlapped reads, and inferring the consensus reads from the layout. String-based methods are derivatives of OLC graph-based methods that attempt to remove duplicate and substring reads before building the graph layouts. The notable string graph algorithms are SGA (Simpson and Durbin 2012) and FALCON (Chaisson et al. 2015), specifically designed to assemble PacBIO long reads. De Bruijn graph is the most widely used *de novo* assembly framework. This approach will divide reads into k-mers representing a node. The overlapping nodes with k-1 bases create an arc in one read, and k-mers that share k-1 bases between the reads construct a direct edge. De Bruijn graph can be classified into Hamiltonian and Eulerian graphs (Conway and Bromage 2011). Hamiltonian kmers represent the nodes, and the edge is the overlap (similar to OLC approach), whereas, in Eulerian method, kmers are the edges. Eulerian approach, used in algorithms such as IDBA-UC (Peng et al. 2012), and SPAdes (Bankevich et al. 2012), is more robust in assembling large genomes than Hamiltonian-based algorithms, such as SOAPdenovo (Luo et al. 2012), velvet (Zerbino and Birney 2008), because it avoids a simplification step required in the construction of the Hamiltonian path (Liao et al. 2019).

Challenges in de Bruijn assembly of a genome include sequencing errors, repetitive regions, and computations resources. These assembly methods assume that the genomic coverage is uniform; however, metagenomic coverage depends on the abundance of that genome in the community. As a result, low abundance genomes in metagenomic sequencing are more likely to end up fragmented. Although algorithms such as Meta-IDBA (Peng et al. 2011), MetaVelvet (Namiki et al. 2012), and metaSPAdes (Nurk et al. 2017), were

built to improve this task, highly fragmented contigs are still common in these assemblies. In chapter two, I will further discuss this problem and present culture-enriched metagenomics, an approach that could potentially address some of these obstacles.

Metagenomic assembly results in thousands of contigs with variable length, but it is unclear where those contigs came from and how many genomes are present in a community. Unsupervised binning of the contigs is a common approach to identifying metagenome assembled genomes (MAGs) (Quince et al. 2017). Binning algorithms predominantly use tetranucleotide frequencies Dick et al. 2009 and coverage information to define similarities across contigs and to cluster them together. The widely used metagenomic binning alogithms include CONCOCT (Alneberg et al. 2014), MetaBAT (Kang et al. 2019), and MaxBin (Wu et al. 2016). Genome-resolved metagenomics allowed the discovery of many microbial groups without culture representative (Brown et al. 2015) and significantly improved microbial genome collections (Nayfach et al. 2019; Xie et al. 2021; Nayfach et al. 2021). However, the metrics that assess the quality of MAGs are not robust. The two metrics that evaluate the quality of MAGs, completeness and contamination based on single-copy core genes, are not sensitive enough and do not assess the quality of accessory genome (Parks et al. 2015; Meyer et al. 2021). In chapter 2, I will compare the length of MAGs with complete whole-genome sequencing (WGS) and discuss how culture-enriched metagenomic can improve the quality of MAGs.

1.6.4 Combination of culture-Independent and -dependent sequencing

With the advancement in DNA sequencing technologies that led to the discovery of new bacterial groups across all taxonomic levels (i.e. new species, genera, families... phyla), the general notion that the human microbiota is not culturable became popular (Rappé and Giovannoni 2003; Stewart 2012) and the human microbiome had been considered unculturable without necessarily testing this hypothesis. In contrast, 48 years ago, Finegold et al. 1974 cultured fecal microbiota of healthy individuals with different diets. They recovered 300 unique species (close to our current estimates of unique bacterial species in a human gut) using a combination of aerobic and anaerobic media conditions. Today, culture-dependent sequencing are at the forefront of innovative microbiome research and the collection of cultured isolates are keep growing (Forster et al. 2019; Poyet et al. 2019; Zou et al. 2019; Aggarwala et al. 2021).

Culture-dependent methods have three main advantages compared to culture-independent sequencing. First, culture distinguishes viable bacteria from dead organisms. Second, selective media conditions allow the growth of the low abundant organisms, often missed by 16S rRNA gene or metagenomic sequencing. And third, building a microbial isolates library for mechanistic and phenotypic investigations. Previous studies from the Surette lab showed that the culture-enrichment increased the number of detected bacterial species of the cystic fibrosis lung microbiota (Sibley et al. 2011; Whelan et al. 2020). Lau et al. 2016 applied culture-enrichment molecular profiling to fecal samples from healthy individuals and IBS patients, and they captured 95% of the OTUs with $> 0.1\%$ relative abundance. It was shown that the majority of microbes captured by culture-independent were recovered by culture; however, culture- dependent profiling identified 3-5 fold more bacterial species (OTUs), suggesting that combining these approaches provides a more comprehensive view of the human microbiota (Lagier et al. 2012; Lau et al. 2016). Whelan et al. 2020 recovered greater taxonomic diversity of the lung microbiota when coupling culture-enrichment with shotgun metagenomics.

In chapters 2 and 3, I will further discuss the advantages of culture-enriched metagenomics and how this approach can provide a higher resolution than culture-independent methods.

1.6.5 Central hypothesis and objectiveness

The importance of the gut microbiome in our health has been well established. As the field moves from microbial associations to microbial treatments in disease, a more in-depth understanding of microbial strains and their functions is necessary. The overarching goal of this thesis is to build bioinformatics tools and approaches to investigate the gut microbiome. I hypothesize that assembly-based metagenomics provides higher resolution than marker-based approaches and that combining culture-enrichment with metagenomics can provide a more comprehensive understanding of intestinal microbiota. This approach will be applied to healthy individuals to capture intestinal microbial diversity. I will also use this method to investigate microbial changes post-FMT to understand the mechanism of microbial engraftment in UC patients.

1.6.6 Aims

To address the above hypothesis, I proposed the following aims:

1. Combine culture-enrichment with shotgun metagenomics to characterize healthy microbiota in eight healthy individuals and compare this approach with culture-independent metagenomics. More specifically, I will investigate whether culture-enriched metagenomics improves the quality of metagenome-assembled genomes and functional annotations (Chapter 2).

2. Conduct culture-enriched metagenomics for a successful FMT donor to compare the microbial composition of UC patients pre- and post-FMT. I will compare 16S rRNA gene amplicon, metagenomics, and culture-enriched metagenomic sequencing to investigate whether these approaches can provide enough resolution to study microbial engraftment. Further, phylogenetic and pangenomic approaches will be applied to examine the mechanism of microbial engraftment in UC patients (Chapter 3).

3. Track longitudinal dynamics of a highly abundant bacteriophage, crAssphage, in a healthy FMT donor. I will investigate whether crAssphage strain from this donor engraft in UC patients post-FMT and compare these dynamics with a publicly available dataset of rCDI patients post-FMT. High-resolution SNP analysis will be applied to the metagenomic samples from UC and rCDI patients to track donor's cAssphage post-FMT (Chapter 4).

4. Compare the gut microbiota of UC patients who received antimicrobial pretreatment before FMT with those who only received FMT in a randomized control trial. 16S rRNA gene amplicon sequencing will be applied to characterize microbial changes in patients compared to donor's microbiota (Chapter 5).

Chapter 2

Culture-enriched metagenomic sequencing of the intestinal microbiota

2.1 Introduction

High-throughput sequencing (HTS) has changed our understanding of the human physiology, particularly the role of the gut microbiome in health and disease. 16S rRNA gene amplicon sequencing has shown the diversity and abundance of human gut microbiota that mainly consist of bacteria but also include bacteriophages, viruses, archaea and fungi. Shotgun metagenomic sequencing provided a higher-resolution view of the complex gut microbiota community and has characterized the intestinal microbial functions in gastrointestinal diseases (e.g. inflammatory bowel disease (Franzosa et al. 2019) and irritable bowel syndrome (Vich Vila et al. 2018)) as well as other systemic disease manifestations (e.g. obesity (Greenblum et al. 2012), Type 2 diabetes (Qin et al. 2012)), and gut-brain axis (Zhu et al. 2020).

Although metagenomic sequencing is the current standard approach to survey microbial taxa and function of the human gut microbiota, this method has a few limitations. First, depending on the source of sampling (e.g. fecal, biopsy, swab, etc), a large proportion of sequenced reads might consist of human DNA (Schmieder and Edwards 2011). Second, low abundant bacterial communities that are an active part of gut microbiota, such as *Enterobactericeae* are poorly covered at typical sequencing depths resulting in few or no sequencing reads (Rajilić-Stojanović and De Vos 2014). As a result, these taxa are often undetected by standard metagenomic pipelines. And third, *de novo* assembly algorithms used to assemble contigs and metagenome assembled genomes (MAGs) from short sequence reads are highly dependent on the coverage information provided by raw metagenomic reads. A single metagenomic sample from a highly complex microbial community can fail to provide sufficient coverage information required to assemble contigs and MAGs accurately (Liao et al. 2019).

Methods for comprehensive culturing of the human gut microbiome have been described (Sibley et al. 2011; Lagier et al. 2012; Rettedal et al. 2014; Lau et al. 2016; Forster et al. 2019; Poyet et al. 2019; Zou et al. 2019; Whelan et al. 2020), which identify greater microbial diversity than culture-independent methods alone. Here, we used culture-enriched metagenomics (CEMG) — shotgun metagenomic sequencing applied to a comprehensive culturing of microbial communities from aerobic and anaerobic media conditions — to characterize the intestinal microbiota of eight healthy individuals, we then compared this approach to shotgun metagenomics, referred here to as direct metagenomics (DMG). We investigated whether CEMG can consistently improve *de novo* assembly of genes and genomes from metagenomic samples across a group of healthy donors. In order to compare these methods, we have established a *de novo* assembly pipeline by benchmarking multiple algorithms.

2.2 Methods

2.2.1 Study design and sample collection

Eight healthy individuals with no gastrointestinal symptoms and no history of antibiotic therapy within three months of the collection were selected for a comprehensive assessment of intestinal microbiota. This project was approved by the Hamilton Integrated Research Ethics Board and conducted at McMaster Children Hospital (Hamilton, ON, Canada).

2.2.2 Culture-enrichment and plate pool libraries

Immediately after defecation, fecal samples were transferred to a sterile container and stored in sealed bags containing an anaerobic pouch (GasPak EZ; BD, MD, USA) and ice-pack. Samples were transferred to the laboratory within three hours of collection and were further processed in an anaerobic chamber (5% CO2, 5% H2, 90% N2; Shel Labs, OR, USA). The sample was cultured using up to 33 media and incubated both anaerobically and aerobically, resulting in 66 culture conditions for culture-enriched molecular profiling using a previously described protocol (Lau et al. 2016). The media and culture conditions were described previously (Lau et al. 2016). 16S rRNA amplicon sequencing was conducted on all 66 culture conditions to determine community composition. To determine a representative subset of culture-enriched plates that adequately represent the sample, the distribution of amplicon sequence variants (ASVs) in the direct sequencing was compared to the culture-enriched sequencing per plate pool using the PLCA algorithm (Whelan et al. 2020). DNA from the plate pools selected using the PLCA algorithm were used for shotgun metagenomics as previously described (Whelan et al. 2020). Supplementary Table A1.1 shows the list of plate pools selected for each sample.

2.2.3 Shotgun metagenomic sequencing

Genomic DNA was extracted using the MagMAX Express 96-Deep Well Magnetic Particle Processor from Applied Biosystems with the Multi-Sample kit (Life Technologies # 4413022) with the addition of a bead beating step. First, samples (0.2g of stool or 300μL of plate pools) were transferred to screw cap tubes containing 2.8mm ceramic beads, 0.1mm glass beads, 100μL of GES (guanidium isothiocyanate, EDTA, N-lauryl sarcosine) and 800μL of 200μM sodium phosphate monobasic, pH8.0. Samples were bead beat at 3000rpm for 3 minutes and centrifuged at 15000rpm. The supernatant was further processed using the Multi-Sample kit. In a 96 well plate, 200uL of the supernatant for each sample was added to 160μL of isopropanol. The plate was sealed and shaken at 505rpm for 3 min. 20 μL of the binding bead mix was added and the plate was shaken again at 505 rpm for 3 min. The plate was then processed on the MagMax express as the manufacturer protocol. The samples were washed twice with wash buffer, then lysis buffer was added, and an RNase treatment was performed. The samples were washed twice again with wash buffer and finally eluted from the beads with two elution buffers in a final volume of 150μL.

DNA concentrations were quantified by Qubit dsDNA HS kit (ThermoFisher Scientific, Mississauga, Canada). Illumina libraries were prepared according to a miniaturized library preparation protocol previously described (Derakhshani et al. 2020), using the NEBNext Ultra II FS DNA Library Prep Kit (NEB, MA, USA). The resulting libraries were subjected to dual size selection using the ProNex Size-Selective Purification System (Promega, WI, USA) to enrich for 800-1000 bp insert sizes. Final libraries were sequenced on an Illumina HiSeq2500 platform in rapid run mode, paired-end 2x250nt, at the McMaster Metagenomics Facility (Hamilton, ON, Canada).

2.2.4 *De novo* assembly and binning

Shotgun metagenomic reads were trimmed using Trimmomatic (Bolger et al. 2014) to remove primer sequences and low quality reads and paired-end libraries interleaved using a custom python script. To build culture-enriched metagenomic assemblies, I have co-assembled trimmed reads from plate pools and fecal samples for each donor using metaSPAde (Bankevich et al. 2012). Trimmed short reads from fecal samples were assembled for each donor separately to build a direct metagenomic (DMG) library.

A custom python script was used to remove contigs $\leq$ 1kb in length. Metabat2 (Kang et al. 2019) was used to assemble metagenomic bins, followed by CheckM to identify metagenome-assembled genomes (MAG). Only bins that contains $\leq$ 10% contamination and $\geq$ 70% completion were defined as MAGs. Contigs that were not present in any bin were defined as unbinned (UnBin). GTDB-tk (Chaumeil et al. 2019) was used for taxonomic classification and multiple sequence alignment of 120 ubiquitous bacterial single-copy proteins. A phylogenetic tree of all MAGs was constructed based on GTDB protein alignment via an approximately-maximum-likelihood model by fasttree (Price et al. 2010). The phylogenetic trees were visualized in R v. 4.0.3. using tidyverse (Wickham et al. 2019), ggtree, ape, ggtreeExtra, and treeio packages.

The cumulative assembly length and total assembly length of contigs $\geq$ 1kb for each sample were calculated in R v. 4.0.3 using the tidyverse package. The most closely related genome of each MAG was identified using GTDB-tk (Chaumeil et al. 2019). The total genomic length of each MAG assembled via DMG and CEMG approaches was compared against their closely related genome in the GTDB whole genome sequence (WGS) library using a log ratio of MAG/WGS. In order to compare the size of MAGs assembled via DMG versus CEMG, microbial families with $\geq$ 5 MAGs were selected and then a linear mixed-effect model was fitted with the sequencing method as the fixed effect as well as microbial families and healthy donors as random effects using lme4 and

lmerTest (Bates et al. 2014) packages in R 4.2.0. Similarly, to compare the assembled size of MAGs within microbial genera that contains $\geq$ 5 MAGs, a linear mixed-effect model was fitted with the sequencing method and genus as fixed effects and healthy donors as random effect. All the figures were visualized in R 4.2.0. using ggplot2 package. All the above scripts are available at https://github.com/SShekarriz/SHCM.

2.2.5 Gene annotation and functional predictions

The genes were annotated in the assembled contigs $\geq$ 1kb using Prokka (Seemann 2014) and the contig ids in Prokka's gff outputs were used to find the position of each gene in Bins, MAGs, and UnBin contigs using a custom code in R 4.2.0. The identified proteins were then clustered at 90% and 70% identity using MMseqs2 (Steinegger and Söding 2017). Next, we used EggNOG-mapper (Cantalapiedra et al. 2021) for functional annotation of all and clustered proteins using cluster of orthologous groups (COGs) (Tatusov et al. 2003), Pfam (Bateman et al. 2004), and the Enzyme Commission (EC) databases. The antimicrobial resistance genes were identified from assembled contigs $\geq$ 1kb using rgi mapper from the CARD database (Alcock et al. 2020). All the data were merged and visualized in R 4.2.0 using the tidyverse (Wickham et al. 2019) package.

2.3 Results

To investigate whether CEMG can enhance *de novo* assembly of gut microbiota contigs and genomes, CEMG and DMG were carried out on fresh fecal samples collected from eight healthy individuals. Briefly, the samples were cultured on up to 66 media conditions, and 16S rRNA gene amplicon sequencing was used to profile the taxonomic composition of each media condition. A subset of plate pools that adequately represent each sample were selected by PLCA (Whelan et al. 2020) for metagenomic sequencing. Supplementary Table A1.1 lists the plates selected for metagenomic sequencing. To compare the *de novo* assembly in CEMG and DMG, we used cumulative assembly length,

quality of MAGs, percentage of genes and functions as *de novo* assembly metrics in the same sample collected from the donors.

2.3.1 Benchmarking of the *de novo* assembly algorithms and methods

In order to construct a robust pipeline for *de novo* assembly of contigs and genomes for metagenomic sequencing, we have benchmarked the performance of multiple assembly and binning algorithms. I compared co-assembly and single sample assembly as well as the performance of two *de novo* prokaryotic assemblers: 1) metaSPAdes (Bankevich et al. 2012), 2) Megahit (Li et al. 2015) and three *de novo* binning softwares: 1) CONCOCT (Alneberg et al. 2014), 2) Maxbin2 (Wu et al. 2015), 3) Metabat2 (Kang et al. 2019). These tools were selected based on the result of a previous study on the critical assessment of metagenomic interpretation (CAMI) (Yue et al. 2020).

A metagenomic sample (B13; supplementary Table A1.1) was assembled using both metaSPAde (Bankevich et al. 2012) and Megahit (Li et al. 2015) to compare the performance of these algorithms. Also to compare co-assembly with single assembly, I used metaSPAde to assemble B13 and B13 + B16 (Table A1.1) samples. These represents two samples from the same donor collected two years apart. To test the performance of binning algorithms, all the assembled samples were binned using Maxbin2 (Wu et al. 2015), CONCOCT (Alneberg et al. 2014) and Metabat2 (Kang et al. 2019). We used MetaQuast (Mikheenko et al. 2015), and CheckM (Parks et al. 2015) to compare the quality of assembly, and binning respectively. Table 2.1 shows the assembly and binning benchmarking result for two DMG samples from a healthy donor.

metaSPADE resolved a longer N50 value — the sequence length of the shortest contig at 50% of the total assembly length — compared to Megahit (Table 2.1). However, Megahit was faster and required less computational memory compared to metaSPADE. These results showed that co-assembly (B13+B16) increased the N50 size compared to

single assembly (B13) from the same healthy individual. These results were consistent with previous findings, highlighting the advantage of co-assembly in improving de Bruijn graph-based assemblies (Sims et al. 2014; Parks et al. 2017). Metabat2 generated the highest number of MAGs, followed by CONCOCT and Maxbin2, for all assembly methods (Table 2.1). Based on the above results and CAMI challenge (Yue et al. 2020), metaSPAde, and metabat2 were used for metagenome assemblies.

TABLE 2.1: Benchmarking multiple assembly and binning softwares.

	metaSPAde	**Megahit**	**metaSPAde**
Sample	**B13**	**B13**	**B13 + B16**
Total Length	378.7 MB	293.7 MB	423.0 MB
Num Contigs	433457	224227	385172
Num Genes (prodigal)	668139	436877	752180
Longest Contig	663.2 kB	343.9 kB	723.4 kB
Shortest Contig	0.1 kB	0.3 kB	1.0 kB
N50	45525	24844	47800
L50	1.3 KB	2.1 KB	8.4 KB
HMM (Campbell et al)	18899	14483	11895
HMM (Rinke et al)	12500	9553	7640
HMM (Ribosomal RNAs)	50	50	153
CONCOCT (bins)	76	73	77
CONCOCT (MAGs)	30	28	34
Maxbin2 (bins)	51	51	53
Maxbin2 (MAGs)	28	27	28
Metabat2 (bins)	83	81	86
Metabat2 (MAGs)	35	35	39

2.3.2 Culture-enriched metagenomics assembles more complete genomic fragments

I applied metaSPAde and Metabat2 to the 8 stool and culture-enriched metagenomic sequencing. This data was used to assess the quality of *de novo* assembly in CEMG compared to DMG. CEMG generated longer contigs compared to the DMG approach (Fig. 2.1A). The total assembly size was more extensive in CEMG than DMG, partly due to the increased sequencing depth. However, the largest contigs in CEMG were bigger than DMG, which resulted in a steeper curve in cumulative assembly length for each donor (Fig. 2.1A). Although the individual's assembly sizes were different, CEMG has shown improved *de novo* assemblies for all healthy donors evaluated (Fig. 2.1A).

In order to investigate the utility of the larger genomic fragments assembled by CEMG, we have binned — grouped contigs that likely originated from the same genome based on coverage information and tetranucleotide frequencies — the contigs per individual and method. We observed that most of the CEMG assembly length was located inside a MAG — a single-taxon bin that has been implicated to be a close representation of an actual individual genome containing $\leq 10\%$ contamination and $\geq 70\%$ completion based on the identified single-copy core genes — across all individuals, while the majority of DMG assemblies were either outside or inside a bin (Fig. 2.1B).

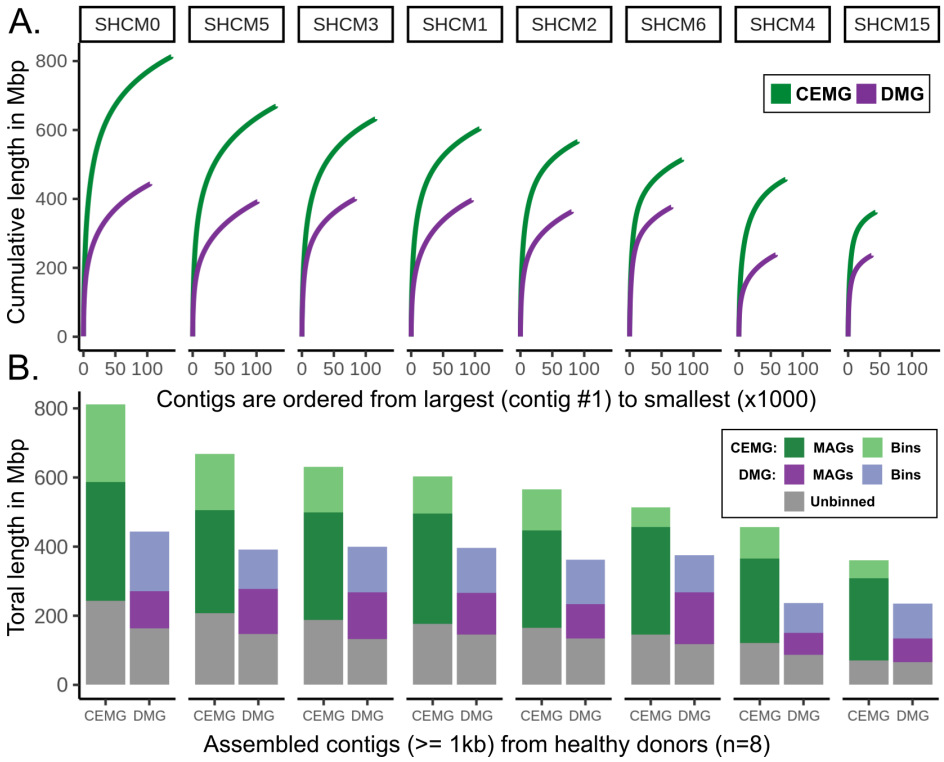

FIGURE 2.1: *De novo* assembled contigs in CEMG vs. DMG. **A.** Cumulative length of assembly for contigs $\geq$ 1kb across eight healthy donors. **B.** Total assembly length of each sample via CEMG and DMG method. Total metagenomic assembly present in CEMG Bin and MAG are shown in light and dark green, respectively. Total metagenomic assembly present in DMG Bin and MAGs are shown in light and dark purple, respectively. The contigs that are not present in a bin are shown in grey (Unbinned).

2.3.3 Culture-enriched metagenomics improves *de novo* assembly of genomes from metagenomics

In order to test whether CEMG systematically improves the assembly of MAGs, the quantity and quality of assembled MAGs in CEMG and DMG was compared to determine whether these methods can equally assemble a diverse groups of bacteria. A phylogenetic tree of MAGs assembled via CEMG and DMG from all the eight donors was constructed. 1255 out of 2823 bins contain the minimum information ($\leq$ 10% contamination and $\geq$ 70% completion) to be reported as MAGs. From the total 1255 MAGs, 879 (%51) and 376 (%34) were generated by CEMG and DMG, respectively. We have used single-copy proteins identified by GTDB to align these MAGs and constructed a phylogenetic tree (Fig. 2.2A). We observed that DMG systematically failed to identify a number of bacterial families such as *Streptococcaceae*, *Enterococcaceae*, *Lactobacillaceae*, *Staphylococcaceae*, and *Bacillaceae* compared to CEMG (Fig. 2.2B). When we mapped DMG raw-reads to these MAGs, we found that they are present in their associated donor sample, but not assembled into a MAG.

Another measure of the quality of MAGs is the total genomic length compared to the expected length. To estimate this, the closest-related genome of each MAG was identified in a publicly available database (GTDB (Chaumeil et al. 2019)) and their total genomic length was compared as a log ratio of MAG/whole-genome sequencing (WGS) (Fig. 2.2C, D). We found that the size of the MAGs assembled via CEMG was significantly (LMM, est=0.05, p=7.7e$-$06) closer to their closely-related genome compared to DMG across the same bacterial families with $\geq$ 5 MAGs (Fig. 2.2C). More specifically, *Staphylococcus, Escherichia, Enterococcus, Streptococcus, Clostridium* and *Flavonifractor* genera with $\geq$ 5 MAGs were not assembled by DMG (Fig. 2.2D). *Eubacterium, Sutterella, Bilophila, Anaerobutyricum, Acetatifactor, Agathobaculum, Faecalibacterium, Dysosmobacter*, and *Gemmiger* genera were significantly different in CEMG than DMG in terms of their genomic size (Fig. 2.2D).

2.3.4 Culture-enriched metagenomics improves gene and functional annotations

Since CEMG resulted in more complete MAGs than DMG, I asked whether these more completed genomes enhance gene and functional predictions. First, the distribution of all predicted genes for each dataset was mapped across MAGs, Bins and unbinned contigs. On average, 50% of genes identified in CEMG across all the samples were present in the MAGs while only 30% of genes in DMG were located inside a MAG (Fig. 2.3A, B). In contrast, the majority of DMG genes (39%) were present in contigs outside of Bins (Fig. 2.3B). As expected, this increases the confidence in assigned genes to specific species/strains

Cluster of Orthologous Groups (COGs) (Tatusov et al. 2003) were used to compare the functional mapping in CEMG and DMG datasets. Although the mean percentage of detected COGs was not significantly different in CEMG and DMG (Anova, CEMG=77.9, DMG=75, se=1.03, p=0.08; Fig. 2.3C), the greatest difference was observed in those COGs that were present in MAGs. On average, 40% of CEMG COGs were identified in MAGs but that number reduced to 24% using the DMG method (Fig. 2.3D).

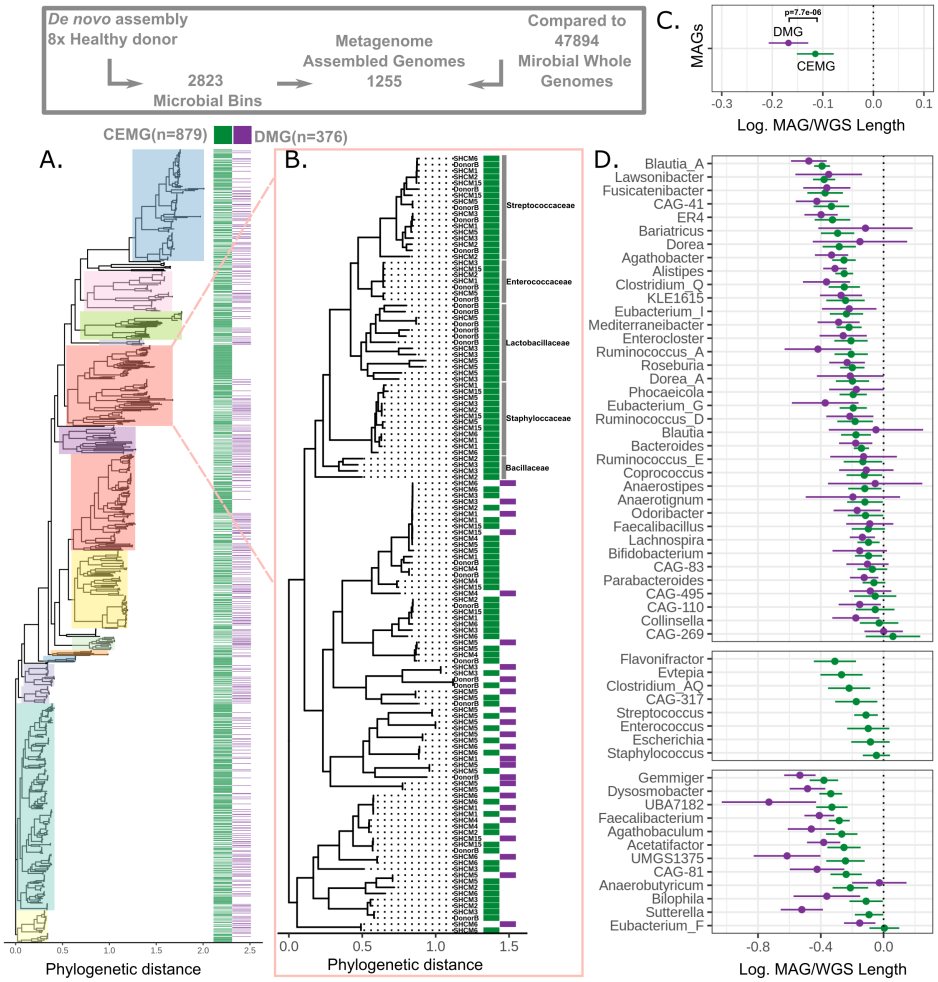

FIGURE 2.2: *De novo* assembled MAGs in CEMG vs. DMG. **A.** A phylogenetic tree of all assembled MAGs from CEMG (n=879) and DMG (n=376) approaches together based on 120 ubiquitous bacterial single-copy proteins alignment. CEMG and DMG MAGs are shown in green and purple, respectively. **B.** An example clade that CEMG MAGs over-represent. **C.** Comparison of the predicted MAGs in CEMG and DMG to their closely related genome in the GTDB database. Each dot shows the mean of MAGs distance compared to a whole-genome sequence (WGS) in GTDB. **D.** Comparison of the predicted MAGs in CEMG and DMG to their closely related genome in GTDB for each genus. The genera only present in CEMG are shown in the middle facet, and those in which their length are significantly different than their closely related genome are shown in the bottom facet.

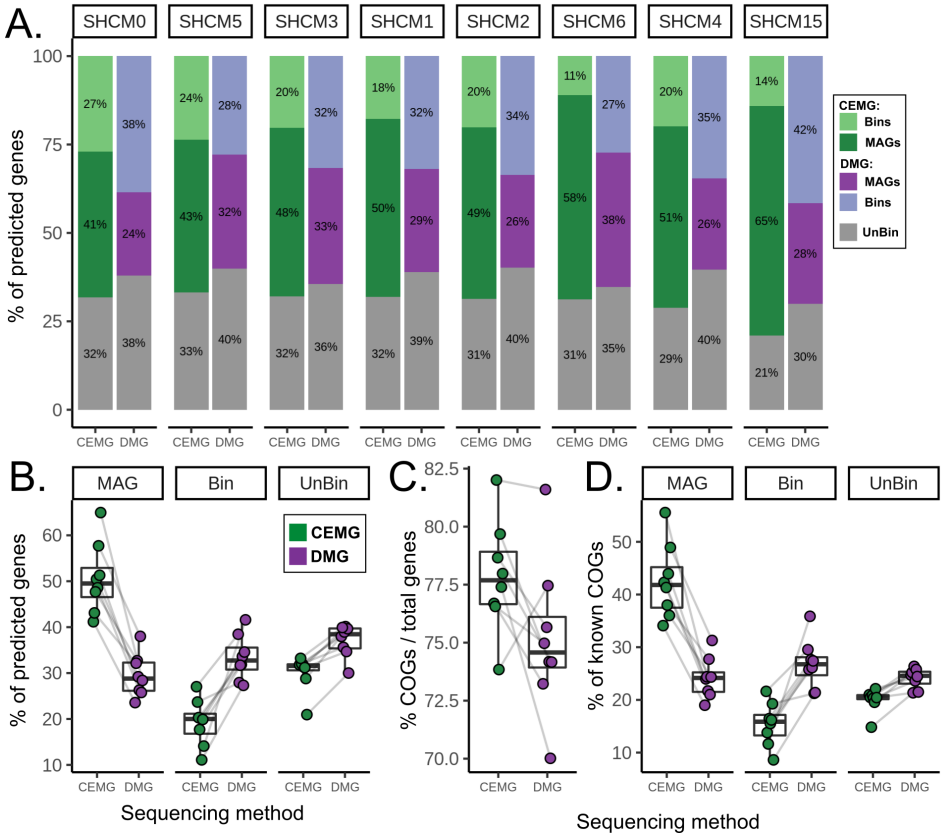

FIGURE 2.3: *De novo* prediction of genes in DMG vs. CEMG. **A.** Percentage of genes present in Bins, MAGs and unBinned contigs in CEMG and DMG methods. **B.** Mean percentage of genes present in MAGs, Bins and UnBinned contigs across eight healthy donors. Mean percentage of genes with known COG functions across metagenomic samples (**C.**) and within MAGs, Bins, and UnBinned contigs (**D.**).

2.3.5 Culture-enriched metagenomics improves detection of antimicrobial resistance genes

To investigate whether CEMG can enhance the detection of antimicrobial resistance (AMR), I used the CARD database (Alcock et al. 2020) to identify AMR genes in CEMG and DMG. No "perfect" AMR hits (proteins with 100% identity to a CARD reference sequence) were identified in MAGs from DMG samples, while on average, 70% of AMR genes from CEMG were located in the MAGs (Fig. 2.4A). Similarly, the mean percentage of strict hits (proteins within the BLAST bit score cut-off) in the MAGs reduced from 50% in CEMG to 25% in DMG. The greatest percentage of perfect and strict AMR hits from DMG were in the contigs outside of any Bin.

In order to test the importance of AMR genes in the genomic context, MAG/Bin taxonomy was used to show the percentage of AMR genes that were identified in each phylum. Interestingly, the greatest percentage of perfect hits that were identified in CEMG MAGs or Bins belonged to *Proteobacteria*, while DMG systematically failed to identify any of these potentially essential hits in all the samples (Fig. 2.4B). A higher percentage of strict hits were observed in *Proteobacteria* in CEMG compared to the DMG method (Fig. 2.4B).

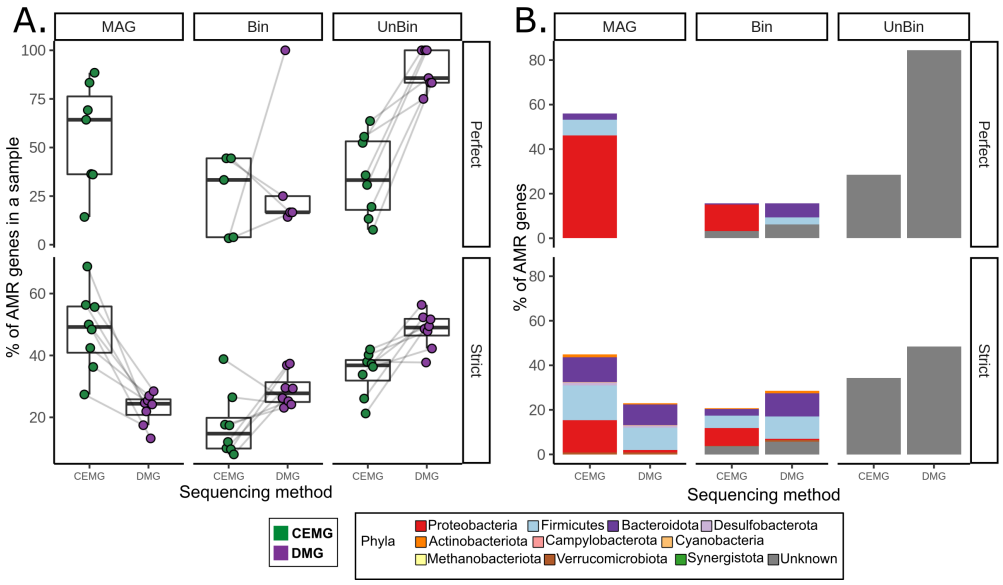

FIGURE 2.4: Antimicrobial resistance genes in DMG vs. CEMG. **A.** Mean percentage of AMR genes present in MAGs, Bins, and UnBinned contigs across CEMG and DMG methods. Each dot shows the percentage of AMR genes detected in healthy donors (n=8). The perfect and strict AMR genes identified by CARD database are shown in top and bottom facets, respectively. **B.** Percentage of AMR genes identified across eight healthy donors. The taxonomy of MAGs and Bins containing AMR genes are shown in color.

2.3.6 Culture-enriched metagenomics predicts more novel proteins.

To determine whether CEMG can discover more novel proteins than DMG, we have clustered the identified proteins in each method at 90% and 70% and annotated these proteins using EggNOG-mapper (Cantalapiedra et al. 2021) across all samples. The total number of proteins increased in CEMG compared to DMG, which was expected given the increased depth of sequencing and greater number of contigs assembled in the CEMG approach (Fig. 2.5A). Further, the mean number of novel proteins — measured as proteins with no close matches in Pfam, EC, and COG databases — was increased in CEMG compared to DMG at 100%, 90%, and 70% clusters (Fig. 2.5B-D). Nevertheless, the differences between these approaches are not significantly different in percentage of proteins with no close match in Pfam, EC, and COG database at 90% clustering threshold (Fig. 2.5E-G), suggesting that CEMG accurately reflects the original metagenomic community. Interestingly, we observed that by increasing the clustering threshold, the percentage of novel proteins has increased, indicating that the unique proteins in the healthy microbiome are more novel than the redundant proteins.

2.4 Discussion

The human microbiota are culturable and it was shown that culture-enriched molecular profiling could provide a more comprehensive view of the microbiota diversity in samples collected from the human intestine (Lagier et al. 2012; Rettedal et al. 2014; Lau et al. 2016), lung (Sibley et al. 2011; Whelan et al. 2020), skin (Myles et al. 2016), and urine (Hilt et al. 2014). We hypothesized that combining culture-enrichment and metagenomic sequencing can improve *de novo* assembly of intestinal microbiota genomes. The simpler microbial communities present in each culture-enriched plate should provide a more even and unique read coverage, which is essential for assembling complex microbial communities in de Bruijn graph-based algorithms (Liao et al. 2019). Furthermore, the distribution of microbes across multiple metagenomic sequencing is expected to improve

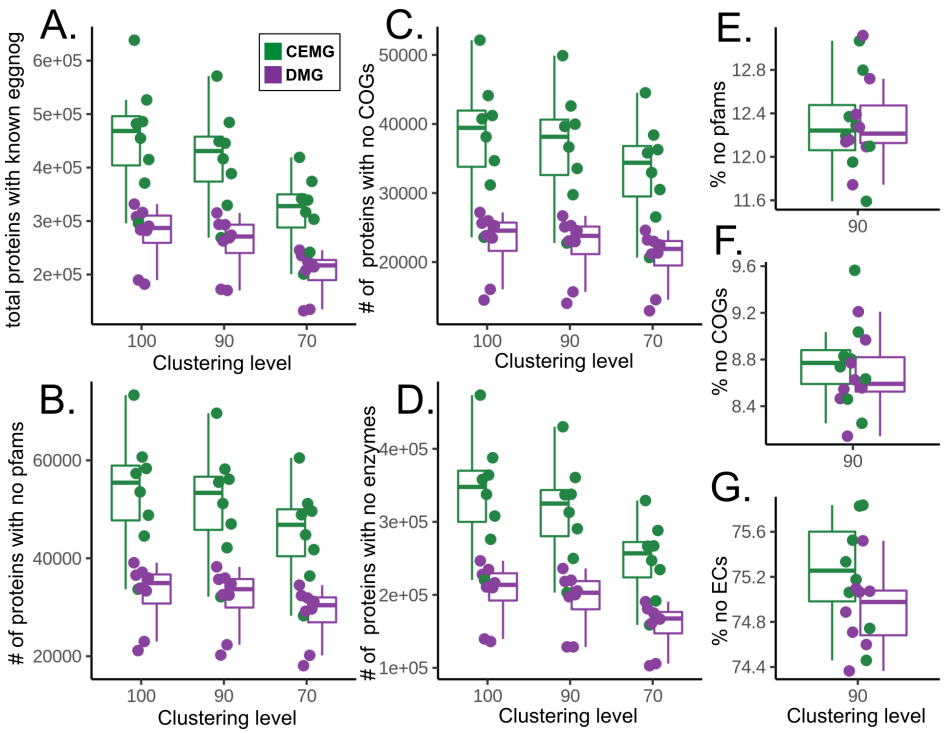

FIGURE 2.5: Prediction of novel proteins in DMG vs. CEMG. **A.** Mean number of predicted proteins at decreasing clustering thresholds across eight healthy donors. Proteins predicted by DMG and CEMG are shown in purple and green, respectively. Mean number of proteins with no Pfams (**B**), COGs assignments (**C**), or Enzyme Commission number (**D**) are shown at different clustering thresholds. Percentage of proteins with no Pfams (**B**), COGs assignments (**C**), or Enzyme Commission number (**D**) are shown at 90%t clustering threshold

binning of assembled contigs. To test this hypothesis, we applied CEMG and DMG to fresh fecal samples collected from eight healthy individuals.

De novo metagenomic assembly is crucial for the future advancement of computational microbiology. It is essential to reconstruct the "meta-genome" structure of the new microbes in an unbiased reference-free approach, given the high heterogeneity in the gut microbiome of healthy individuals (Lozupone et al. 2012) and incomplete microbial reference databases (Loeffler et al. 2020). De Bruijn graph-based methods predominantly used by *de novo* metagenomic assembler are impaired by sequencing errors, genomic repeats, and uneven sequencing coverage information. Our results have shown that

CEMG could improve de Bruijn graphs by providing unique coverage information that help these algorithms to find matched k-mers across assembly nodes and fill gaps across the assembly scaffolds.

Although there are multiple large-scale benchmarking studies to assess metagenomic interpretations (Yue et al. 2020; Sczyrba et al. 2017; Meyer et al. 2021), it was shown that the performance of these algorithms depends on the microbial community's complexity and the depth of metagenomic sequencing (Fritz et al. 2019). Widely used algorithms for *de novo* assembly and binning of metagenomic contigs were compared based on the reference-independent (N50) and -dependent (single-core genes) metrics. Our results indicated that Eulerian de Bruijn algorithms (such as SPAdes) resolve longer contigs compared to other memory efficient de Bruijn methods (such as MEGAHIT), presumably because they do not partition reads using k-mer abundance patterns. However, SPAdes required a memory-intensive machine and a longer run time. The binning algorithms showed variable results, and the number of identified MAGs was sample specific. However, our results indicated that CEMG reduced the variability of binning algorithms indicating that unique coverage information provided by plate pools improved the binning of metagenomic contigs.

Genome-resolved metagenomics, the construction of MAGs, has helped us to further understand the diversity and functions of microbial strains in healthy human gut microbiota (Almeida et al. 2019). However, these MAGs are often not accurate representations of the bacterial genomes. A common challenge with the assembly of MAGs is over or under estimation of genomic size. The standards metrics that assess MAG quality (the number of single-copy core genes) are not necessarily robust enough to report this issue. We found that MAGs resolved by CEMG were significantly more similar to the closest complete bacterial genomes than DMG. Interestingly, some bacterial families that were present in DMG as mapped reads, were only assembled as MAGs in the CEMG method.

The larger genome fragments assembled by CEMG have improved the gene and functional annotations compared to DMG because they reduce the chance of missing open reading frames via gene prediction tools such as Prodigal (Hyatt et al. 2010). We also observed the greatest percentage of genes, functions, and AMR genes in the CEMG MAGs. Detection of AMR genes in a MAG is crucial because it provides contextualized information about the resistance mechanism. The DMG method has failed to identify any perfect *Proteobacteria sp.* AMR hits in a MAG, implicating the importance of culture-based methods to identify these high-priority resistance bacteria in a sample. The total proportions of COGs and proteins clustered at multiple thresholds were not significantly different between these two methods, suggesting that CEMG is not biased towards any bacterial groups and could show an accurate representation of original bacterial communities present in DMG.

CEMG is not a replacement for DMG, instead it is an approach that in combination with DMG can significantly enhance *de novo* assembly of microbial genes, functions, and genomes from metagenomics. This approach is labour-intensive and possibly not feasible for large sample size studies but can be used to build a comprehensive assembly for key reference samples in addition to providing a strain collection library. In chapter 3, I used this approach for high resolution characterization of a healthy donor to investigate the mechanism of microbial engraftment after fecal microbiota transplantation.

Chapter 3

Culture-enriched metagenomics reveals microbial engraftment after FMT in patients with ulcerative colitis

3.1 Introduction

ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD) restricted to the colon and of unknown etiology (Kappelman et al. 2007). UC is generally considered to arise due to a disruption in the balance between the immune system and microbiota in a genetically susceptible individual (De Souza and Fiocchi 2016; Hindryckx et al. 2016). Current standard medical treatments have focused on supressing the immune response and are not always effective at controlling disease (Talley et al. 2011). An alternative approach is to alter the microbial environment responsible for driving the immune response (Moayyedi 2018). fecal microbiota transplantation (FMT) has emerged as an increasingly popular approach to alter the colonic microbiota (Fuentes et al. 2017)

and is a standard therapy for patients with recurrent- *Clostridioides difficile* infection (rCDI) (Khoruts et al. 2021; Khanna et al. 2017a) and has also been evaluated in UC (Moayyedi et al. 2015; Rossen et al. 2015; Costello et al. 2019; Paramsothy et al. 2017; Haifer et al. 2022b).

Moayyedi et al. (2015) reported on the first randomized controlled trial (RCT) of FMT for patients with active UC. This RCT showed that the percentage of patients with active UC in which remission was induced after FMT (24%) was significantly higher than the placebo (5%), with no difference in adverse events. This has been replicated by other researchers and there are now four RCTs suggesting FMT is efficacious in UC (Narula et al. 2017; Paramsothy et al. 2017; Costello et al. 2019; Haifer et al. 2022b). One of the donors (donor B) involved in the trial reported by Moayyedi et al. (2015), was more successful compared to other donors, with 7 of the 9 responders — defined as a Mayo score <3 and complete healing of the mucosa at flexible sigmoidoscopy at week 7 — in the trial receiving FMT from donor B. This apparent donor effect cannot be studied in some trials as they used a mixture of donors rather than narrowing the pool of donors and only using one donor per patient. In this study, we built upon the RCT of Moayyedi et al. (2015) by further investigating the microbial composition of patients who received FMT from donor B compared to those who received placebo treatments to ask whether a specific group of microbes were engrafted following FMT and to determine whether microbial engraftment is associated with remission post-FMT. Previous studies have characterised microbial enrichments — an increase in the relative abundance of observed bacteria — following FMT using 16S rRNA gene amplicon (rCDI, (Weingarden et al. 2015; Khanna et al. 2017b)) and marker-based metagenomics (UC, (Paramsothy et al. 2019)). However, microbial engraftment — the transfer of microbes from donor to patients — following FMT has yet to be determined, especially given the low strain/species resolution provided by 16S rRNA gene amplicon. These culture-independent approaches are often not sensitive enough to capture low-abundant bacteria

(Lau et al. 2016). The bacteria identified via 16S rRNA gene amplicon or metagenomic sequencing following FMT may be present before FMT but below the detection level for culture-independent methods. The culture-enriched sequencing methods provide a more comprehensive view of the human microbiome than culture-independent sequencing, particularly for low abundant bacteria, and past studies showed the utility of this approach to capture the diversity of intestinal (Lagier et al. 2012; Rettedal et al. 2014; Lau et al. 2016), lung (Sibley et al. 2011; Whelan et al. 2020), skin (Myles et al. 2016), and urine (Hilt et al. 2014) human microbiota.

To answer the question of whether specific groups of microbes are responsible for inducing remission in UC, we have therefore used three high-throughput sequencing approaches; 16S rRNA gene amplicon, metagenomics and culture-enriched metagenomics (CEMG). Further, we asked whether the observed enrichment post-FMT was due to the patients' own microbiome being restored through FMT (e.g., the enrichment of low abundance bacteria that were originally below the level of detection) or due to engraftment of the organisms from the donor.

3.2 Methods

3.2.1 Study design and sample collection

The study design and sample collection as described earlier (Moayyedi et al. 2015). Briefly, 70 active UC patients (Mayo score $\geq$4 with an endoscopic Mayo score $\geq$1) randomly assigned to either 6 weeks of FMT (once per week; 50 mL, via enema, from healthy anonymous donor) or placebo (once per week; 50 mL water enema) in a double-blind randomized controlled trial. The stool samples were collected at baseline, before the FMT, and during each week of FMT.

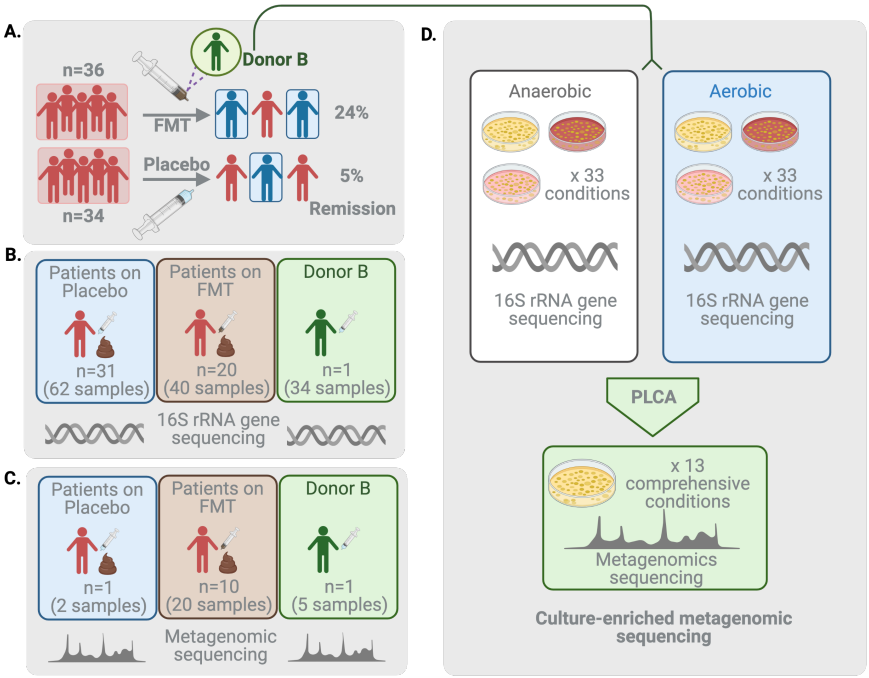

FIGURE 3.1: **Graphical illustration of the methodology used in this study A**. Moayyedi et al. (2015) double blind randomized control trial of FMT for UC patients. Donor B (**green**) was more successful compared to the five other donors involved in the trial. **B**. 16S rRNA gene amplicon sequencing was conducted for all the patients who received FMT from donor B or placebo. **C**. Shotgun metagenomic sequencing was carried out for a subset of patients who received FMT from donor B. **D**. The culture-enriched metagenomics workflow to build a comprehensive microbiome database of donor B's gut microbiome.

3.2.2 DNA extraction and 16S rRNA gene sequencing

Genomic DNA extraction and PCR amplification of the V3 region of 16S rRNA gene, was conducted using previously described protocols (Whelan et al. 2014; Bartram et al. 2011; Moayyedi et al. 2015). Briefly, 0.2 g of fecal matter was mechanically homogenized using ceramic beads in 800 μL of 200 mM NaPO 4 (pH 8) and 100 μL of guanidine thiocyanate-EDTA-N-lauroyl sacosine. This was followed by enzymatic lysis of the supernatant using 50 μL of 100 mg/mL lysozyme, 50 μL of 10 U/μL mutanolysin, and 10 μL of 10 mg/mL RNase A for one hour at 37 °C. Then, 25 μL of 25% sodium dodecyl sulfate (SDS), 25 μL of 20 mg/mL proteinase K, and 75 μL of 5 M NaCl was added, and incubated for one

hour at 65°C. Supernatants were collected and purified through the addition of phenol-chloroform-isoamyl alcohol (25:24:1; Sigma, St. Louis, MO, USA). DNA was recovered using the DNA Clean & Concentrator TM -25 columns, as per manufacturer's instructions (Zymo, Irvine, CA, USA) and quantified using the NanoDrop (Thermofisher, Burlington, ON). After genomic DNA extraction, the V3 region of the 16S rRNA gene was amplified via PCR using these conditions per reaction well: Total polymerase chain reaction volume of 50 μL (5 μL of 10X buffer, 1.5 μL of 50mM MgCl 2 , 1 μL of 10 mM dNTPs, 2 μL of 10mg/mL BSA, 5 μL of 1 μM of each primer, 0.25 μL of Taq polymerase (1.25U/ μL), and 30.25 μL of dH 2 O. Each reaction was divided into triplicate for greater efficiency. The primers used in this study were developed by Bartram et al.,2011. PCR conditions used included an initial denaturation at 94°C for 2 minutes, followed by 30 cycles of 94°C for 30s, 50°C for 30s, 72°C for 30s, followed by a final elongation at 72°C for 10 minutes. All samples were sequenced using an Ilumina MiSeq platform at the McMaster Genome Facility (Hamilton, Ontario, Canada). Samples were processed in batches, meaning not all samples were extracted and sequenced at the same time.

3.2.3 16S rRNA gene sequencing processing pipeline

Cutadapt v. 1.14 (Martin 2011) was used to filter and trim adapter sequences and PCR primers from the raw reads, using a quality score cut-off of 30 and a minimum read length of 100 bp. We used DADA2 (Callahan et al. 2016) to resolve the sequence variants from the trimmed raw reads as follow. DNA sequences were trimmed and filtered based on the quality of the reads for each Illumina run separately. The Illumina sequencing error rates were detected, and sequences were denosied to produce ASV count table. The sequence variant tables from the different Illumina runs were merged to produce a single ASV table. Chimeras were removed and taxonomy was assigned using the DADA2 implementation of the RDP classifier against the SILVA database v. 1.3.2 (Quast et al. 2012), at 50% bootstrap confidence. All downstream analysis was conducted in R v. 4.0.3

(Ihaka and Gentleman 1996). We curated the data and generated plots using phyloseq v. 1.22.3 (McMurdie and Holmes 2013) and the following tidyverse packages: dplyr v. 0.7.6, tidyr v. 0.8.1, rlang v. 0.2.1, and ggplot2 v. 3.0.0. To visualize sample distances (beta-diversity), we calculated both Aitchison and Bray–Curtis distances. ASV counts transformed to the centered log-ratio (CLR) using microbiome v.1.12.0 and visualized via Principal Component Analysis (PCA) for Aitchison distances. We applied PCoA to generate Bray-Curtis distances for ordination plots and unweighted pair group method with arithmetic mean (UPGMA) for clustering trees using ape package v. 5.2 (Paradis et al. 2004) and hclust() function in R. Trees were visualized using the stringi package v. 1.2.3 in R and the Interactive Tree of Life (iTOL) (Letunic and Bork 2007). To measure sample diversity, we calculated Shannon values for sample or group using phyloseq package and visualize them using ggplot. The variability of microbiota was tested by PERMANOVA on Bray-curtis distances based on relative-abundance of microbes in each sample using adonis() function in ape package (Paradis et al. 2004). The diversity of samples calculated by Shannon values using phyloseq and the significant changes were measured by Linear Mixed-Effects Models using lmer package. Those ASVs that were present in ≥ 1 sample from donor B selected as donor B's ASVs. Then, the donor B's ASVs were compared in each patient with data from prior and post-FMT, ASVs with a relative abundance of 0 in a patient before FMT and $\geq 0.1\%$ post-FMT were labelled as engrafted. In order to find the commonly engrafted ASVs, the number of engrafted ASVs was compared across an increasing number of patients in FMT vs. placebo group. To have an equal number of patients across these two groups, 20 of 31 patients on placebo treatment were randomly sampled (with 100 re-sampling).

3.2.4 Library preparation and read-based shotgun metagenomics pipeline

We have conducted shotgun metagenomics on 22 samples collected from 11 patients, with 2-time points each, in this study (4 non-responder, 6 responders patients who received FMT and one patient on placebo). Genomic DNA was standardized to 5 ng/ μL and sonicated to 500 bp. Using the NEBNext Multiplex Oligos for Illumina kit (New England Biolabs), DNA ends were blunted, adapter ligated, PCR amplified, and cleaned as per manufacturers instructions. Library preparations were sent to the McMaster Genome Facility, and sequenced using the Illumina HiSeq platform. The forward and reverse sequencing runs were concatanated and trimmed for primer adapters and low quality reads using Trimmomatic (Bolger et al. 2014). The taxonomic, and gene-family composition of trimmed shotgun reads identified using Metaphlan2 and Humann2 pipeline (Franzosa et al. 2018). All downstream analysis was conducted in R v. 4.0.3. The Bray-Curtis distances calculated based on the relative abundance of known species and gene-families using phyloseq package. Principal coordinate analysis (PCoA) plots were generated using phyloseq and ggplot2. Unweighted pair group method with arithmetic mean (UPGMA) trees based on Bray–Curtis distances were generated using the ape package v. 5.2 (Paradis et al. 2004) and hclust() function in R. Trees were visualized using the stringi package v. 1.2.3 in R and the Interactive Tree of Life (iTOL). (Letunic and Bork 2007). The diversity of samples measured by Shannon index using phyloseq and the significant changes were measured by Linear Mixed-Effects Models using lmer package. For the microbial taxonomy dataset, the engrafted strains were defined as any strains present in ≥ 1 sample from donor B with relative abundance of 0 prior to FMT and 0.1% post-FMT in a patient. Humann2 uses a detection threshold of 0.01% relative abundance which is equivalent to 0.1x fold-coverage of a 5 Mbp microbial genome (Franzosa et al. 2018). Given 1000 genes per Mbp, we expect 0.0005% relative abundance for detection of a gene family. Thus, any gene family with a minimum relative abundance

of 0.0005% in donor B samples, 0% before FMT and 0.0005% post- FMT defined as engrafted gene-families.

3.2.5 Culture-enriched and independent metagenomics on donor B samples

A single fresh, anaerobic fecal sample collected from donor B. The collected sample was cultured using 33 media, and incubation of plates anaerobically and aerobically resulted in 66 culture conditions for culture-enriched molecular profiling using previously described protocol (Lau et al. 2016). The list of media and culture conditions are described earlier (Lau et al. 2016). 16Sr RNA amplicon sequencing was conducted on all the 66 culture conditions. To determine a representative subset of culture-enriched plates that adequately represent the sample, the distribution of ASVs in the direct sequencing was compared to the culture-enriched sequencing per plate pool using the PLCA algorithm (Whelan et al. 2020). Shotgun metagenomics was conducted on the subset of plate pools identified by the PLCA algorithm. Genomic DNA was isolated from the thirteen selected plate pool and shotgun metagenomics conducted as previously described (Whelan et al. 2020; Lau et al. 2016). Direct shotgun metagenomics conducted on the same fecal sample, which was earlier used for culture-enriched metagenomics as well as three other fecal samples collected from donor B at different time points (2013, 2017x2).

3.2.6 Comparison of the culture-enriched metagenomics with direct metagenomics data

To build the culture-enriched metagenomic library, the raw shotgun sequences from the selected plate pools and the original fecal sample collected from donor B co-assembled

together as follows. The low-quality reads and sequencing primers removed using Trimmomatic (Bolger et al. 2014). The reads decontaminated for any human DNA utilizing DeconSeq package (Schmieder and Edwards 2011). The shotgun reads were co-assembled and binned using metaSPADE (Bankevich et al. 2012) and Metabat2 (Kang et al. 2019) respectively. In addition to CEMG assembly, the microbial composition of direct metagenomics (DMG) from the fecal sample assembled and binned separately. These two datasets are labelled as CEMG and DMG in Figure A2.3. The microbial composition of DMG and CEMG datasets were then comprehensively evaluated using the following procedure. The single-copy core genes were identified within each bin using CheckM (Parks et al. 2015), any bin with minimum 70% completion and maximum 10% contamination were defined as a metagenome assembled genome (MAG). The shotgun reads were mapped to the assembled contigs to estimate sequence coverage for all contigs, Bins, MAGs, and those contigs that were not present in any bin. We used bwa (Li and Durbin 2009) to map reads to assembled contigs and anvio pipeline (Eren et al. 2015) to normalize the coverage to the depth of sequencing. The detection values calculated for each bin using anvio package (Eren et al. 2015). The detection value defined as the proportion of a given MAG that is covered at least 1X; in other words, it estimates the proportion of MAG that recruited reads to it. We used GTDB-Tk (Chaumeil et al. 2019) to build a phylogenetic tree, and taxonomic assignment of MAGs. All of the figures visualized in R v. 4.0.3.

3.2.7 Microbial engraftment detection in metagenomic data

To detect microbial engraftment, we aimed to construct a comprehensive library of microbial genes and genomes from donor B. This library contains 4 DMG samples and single CEMG sequencing methods. The low-quality reads and sequencing primers removed using Trimmomatic (Bolger et al. 2014). The reads decontaminated for any human DNA utilizing DeconSeq package (Schmieder and Edwards 2011). The shotgun

reads from both culture-dependent and independent libraries were co-assembled and binned using metaSPADE (Bankevich et al. 2012) and Metabat2 (Kang et al. 2019) respectively. The MAGs and MABs were identified using previously described criteria. The single-copy core genes were identified within each bin using CheckM (Parks et al. 2015), any bin with minimum 50% completion and maximum 10% contamination were defined as a metagenome assembled genome (MAG). To include more number of MAGs in our database, we reduced the completion value of a MAG from 70% to 50%. We used Prodigal (Hyatt et al. 2010) to predict prokaryotic genes and coding DNA sequences (CDS) from the assembled contigs. The taxonomic labels are assigned to all bins using GTDB-Tk (Chaumeil et al. 2019). In total, we were able to assemble 255 metagenomics assembled genomes (minimum 50% completion and maximum 10% contamination) and 1,130,000 completed prokaryotic genes. After de-novo prediction of genes and MAGs, we mapped the collected metagenomics samples from before and after FMT (22 samples in total) to the assembled contigs from donor B. The raw reads from each sample were mapped to the assembled contigs using bwa mem (Li and Durbin 2009) and the coverage information normalized to the depth of sequencing using anvio package. For each MAG, the detection and single nucleotide variability measurements calculated using anvio pipeline. The variability index shows the number of reported single-nucleotide variants per kilobase pair. All the downstream analyses to detect microbial engraftment at gene and genomic-level were performed in R v. 4.0.3 R.

The assembled MAGs from donor B were classified by comparing the short read mapping coverage and SNV frequencies from before and after FMT for each patient. **Shared** category was defined as MAGs covered above our minimum detection cutoff ($\geq$ 60% proportion of nucleotides in a MAG that has at least 1X coverage) in a patient both before and after FMT. The MAGs with coverage lower than the minimum detection cutoff in both time points were classified as **Unique to Donor**. The MAGs with 0 coverage before FMT and $\geq$ 60% post-FMT were classified as **Engrafted** and opposite

cutoffs were used for **Lost** category. To measure variability across detected MAGs, the SNV frequency calculated for MAGs with minimum 0.6% coverage in both time points using anvio pipeline (Eren et al. 2015). The SNV frequency shows the number of single-nucleotide variants per kilo base pair. The **Shared** MAGs that showed >=1 SNV per kilobase pair before FMT but their frequencies reduced to <= 0.5 per kbp after FMT were classified as **Strain Replacement**. In other words, those MAGs that were present in both time points but highly similar to donor B's original MAG only after FMT are defined as strain replacement.

To detect microbial engraftment at the gene-level, we compared the coverage of all the 1,130,000 donor B microbial genes across UC patients before and after FMT. In this model, those genes that their detection (% of gene covered at least 1X) was 0 before FMT and became at least 0.6 with minimum 5X coverage after FMT is called engrafted genes (Fig. 3.5 C). To narrow down the number of engrafted genes, commonly engrafted ones across three patients were labelled as common engraftment. This model applied to all eleven patients, regardless of their response to FMT. Then, we compared these commonly engrafted genes against the Uniref90 reference database using Diamond blastp and the identified Uniprot ids annotated with GO, KEGG, COG, Pfams, and lineage information.

3.2.8 Single whole-genome sequencing and comparative genomics

30 *Dorea*, 1 *F.prausnitzii*, and 67 *Blautia* strains were isolated from human gut. The media, culture conditions for isolation, library preparation, and sequencing protocols as described earlier (Derakhshani et al. 2020). In addition, we have collected 65 *Dorea*, 98 *F.prausnitzii*, and 143 *Blautia* strains available in NCBI RefSeq (May 2020). We have annotated all the genes and CDS using Prokka (Seemann 2014) with default settings. The assembled genomes were re-classified using GTDB-Tk (Chaumeil et al. 2019) and

phylogenetic trees were constructed within each genus based on multiple sequence alignment of 120 bacterial marker genes from GTDB database (Parks et al. 2020). We used panaroo (Tonkin-Hill et al. 2020) with strict mode and mafft aligner to generate core-gene alignment within each species. We then used approximately-maximum-likelihood model via FastTree (Price et al. 2010) to construct phylogenetic trees for strains within each species. We made a blastn database using all the 402 genomes and tracked the commonly engrafted genes across these genomes with a minimum ≥90 pident and qcovhsp ≥90 cut-offs. The number of non-redundant positive hits from blastn output were visualized for each genome on the phylogenetic trees for genus and species collections. All the phylogenetic trees were visualized in v. 1.2.3 R using ggtree, ggtreeExtra, and ape packages.

A single genome with the most number of commonly engrafted genes and fewest contigs were selected for *D. longicatena*, *F. prausnitzii*, and *F. saccharivorans* as representative strains of commonly engrafted genes. Then we mapped all the shotgun raw-reads from donor B (5 samples) and UC patients (22 samples) to these three genomes using bwa-mem (Li and Durbin 2009). The commonly engrafted genes identified for each representative strain using previously described gene engraftment model. Briefly, genes that were not present (0% 1X coverage) pre-FMT but present (>0.6% with ≥5X coverage) post-FMT across ≥3 patients were selected as commonly engrafted genes. We used a custom python code to extract all the commonly engrafted genes and their flanking regions (20,000 bps) from the three representative strains. To find whether the engrafted genes are the result HGT or strain replacement, we used anvio pipeline (Eren et al. 2015) for de-novo characterization and reporting of SNVs for the two selected flanking regions in *F.prausnitzii* and *F. saccharivorans* strains. In short, a table of nucleotide base frequencies for the 80,000 bp gene clusters, contained commonly engrafted genes, were constructed for *F. saccharivorans* and *F.prausnitzii* representative strains. The consensus nucleotide identified based on anvio's conservative heuristic model. We then

selected and visualized only those base positions that were identical across all donor B samples (5 samples). These are donor B's specific SNVs that we used to see whether the samples collected after FMT are closer to the donor's SNV profile (less number of SNV) or contain increased SNVs. The SNV tables were filtered and visualized in v. 1.2.3 in R using tidyverse packages.

3.2.9 Species- and strain-specific markers for a metagenomic survey of IBD patients and healthy controls

To build species-specific marker, a group of genomes from distinct species were selected from *Dorea sp.*, *Feacalibacterium sp.*, and *Fusicatenibacter sp.* collections for pangenome analysis. Species-specific core-genomes were identified and visualized using Anvio microbial pangenomics workflow. Brierly, gene calls were annotated with prodigal and MCL algorithm (Van Dongen and Abreu-Goodger 2012) was used to identify gene-cluster across the pangenome alignment with –mcl-inflation 10 –minbit 0.5 –use-ncbi-blast. To test the accuracy of species-specific and strain-specific (the commonly engrafted genes) markers, we have used 1200 WGS from our lab strain collections. These strains are diverse bacterial isolates from all bacterial phyla collected from the human microbiota. We mapped shotgun reads from 1112 WGS to markers using bwa-mem (Li and Durbin 2009) with –B 40 –O 60 –E 10 —L 100 parameters to find perfectly aligned reads over their entire length and samtools (Li et al. 2009) to extract coverage information from bam file. We then used the percentage of 1X coverage to visualize the coverage information for each marker in v. 1.2.3 in R using tidyverse packages. We used a publicly available metagenomic dataset (PRJNA279196 (Franzosa et al. 2019)) to investigate the specificity of *D. longicatena*, *F. prausnitzii*, and *F. saccharivorans* strains in IBD patients compared to healthy controls. We downloaded metagenomic samples from the SRA database via Entrez Direct (EDirect) command line. Metagenomic shotgun reads from all the samples (n=220) were mapped to marker gene-clusters using bwa-mem (Li

and Durbin 2009) with the parameters specified above and samtools (Li et al. 2009). Subsequently, the percentage of 1X coverage of strain- and species-specific markers were visualized in v. 1.2.3 in R using tidyverse packages.

3.3 Results

We collected samples from 51 patients (paired samples before and after FMT) who randomly assigned to 6 weeks of FMT from donor B (n=20) or placebo treatment (n=31) once per week (Moayyedi et al. 2015) as well as 34 fecal slurries from donor B collected during the clinical trial (Fig. 3.1A). We sequenced the variable 3 region of the 16S rRNA gene amplicon from all samples (Fig. 3.1B) and conducted shotgun metagenomics for ten patients who received FMT from donor B (six responders and four non-responders to FMT), one patient who randomized to the placebo group, and five samples from donor B (Fig. 3.1C). We built a comprehensive *de novo* sequence library of donor B via culture-enriched metagenomics. 16S rRNA gene amplicon sequencing was carried out on 66 growth conditions (33 anaerobic and 33 aerobic, see Methods), and metagenomic sequencing on the 13 most comprehensive plate pools selected using a plate coverage algorithm, as previously described (Whelan et al. 2020) (Fig. 3.1D).

3.3.1 16S rRNA gene sequencing does not provide the necessary resolution to determine if engraftment is occurring

We reanalysed the 16S rRNA gene sequencing data of donor B fecal slurries as well as patients who received FMT from donor B or placebo from Moayyedi et al. (2015), using higher resolution amplicon sequence variants (ASV) (Callahan et al. 2016). We hypothesised that remission induced by FMT would be associated with changes in the microbial composition of patients before and after FMT, either due to enrichment (increase in relative abundance of ASVs present at baseline) or engraftment (detection of donor-specific ASVs following FMT). To do so, we compared paired patient samples

collected prior to and six weeks post-treatment. The 16S rRNA gene dataset contained 102 samples from 51 patients and 34 donor B samples (Fig. 3.1B). This dataset includes 8 patients who went into remission following FMT (6 who received FMT, and two from the placebo group).

We calculated the community-wide distance between each sample using the Bray-Curtis beta-diversity metric and clustered all samples into a UPGMA tree (Fig. 3.2A) FMT recipients were less likely to cluster most closely to their pre-treatment sample compared to patients given placebo (45% vs. 71%, respectively) indicating smaller community change in the placebo compared to the FMT group, as previously describe (Moayyedi et al. 2015). Further, there was no significant separation based on FMT or placebo treatment regardless of whether all patients were analyzed or samples were sub-setted to just responder or non-responder (ANOVA non-parametric test based on Bray-Curtis distance; FMT R2 =0.02, p-value=0.2; Placebo R2 =0.01, p-value=0.06; Responder R2=0.06, p-value=0.2; Non-responder R2=0.02, p-value=0.3 Fig. 3.2B; or based on Aitchison distance Supplementary Figure A2.1B). The alpha diversity of the microbiota, as measured by the Shannon index, did not significantly change in week six following either FMT (linear mixed-effect model, p-value=0.8) or placebo treatment (linear mixed-effect model, p-value=0.5) when compared to baseline. These findings suggest that there is not a common change to the microbial community post-FMT in UC patients, most likely due to the heterogeneity in the microbiome composition across UC patients (R2=0.71, p-value=0.001). However, the microbial composition of samples collected from donor B is significantly different from the rest of the samples in this cohort (R2 =0.08, p-value=0.001; Fig. 3.2B, and Supplementary Figure A2.1B). Using 16S rRNA gene sequencing data, we were not able to find a global difference between samples collected from patients before and after FMT or an association between remission and microbial composition.

To test whether microbial engraftment is associated with remission post-FMT, we determined any ASVs from donor B that could be potentially engrafted based on their relative abundance in each patient (donor B ASVs with a relative abundance of 0 in a patient before FMT and $\geq 0.1\%$ post-FMT). If we used this definition of engraftment, on average, 164 and 63 donor B's ASVs were engrafted ($\geq$1 patient) in the FMT and placebo groups, respectively (Fig. 3.2C). These results suggest that there is a moderately high false positive in detecting true engraftment via 16S rRNA gene amplicon data even with the stringent cut-offs used here. Although there is a signature of 25 ASVs commonly engrafted in $\geq$ 3 patients post-FMT (Fig. 3.2C), these ASVs are not unique to those who responded to FMT treatment (Supplementary Figure A2.1D). We conclude that low resolution provided by 16S rRNA gene sequencing results in too high of an error rate to accurately predict microbial engraftment.

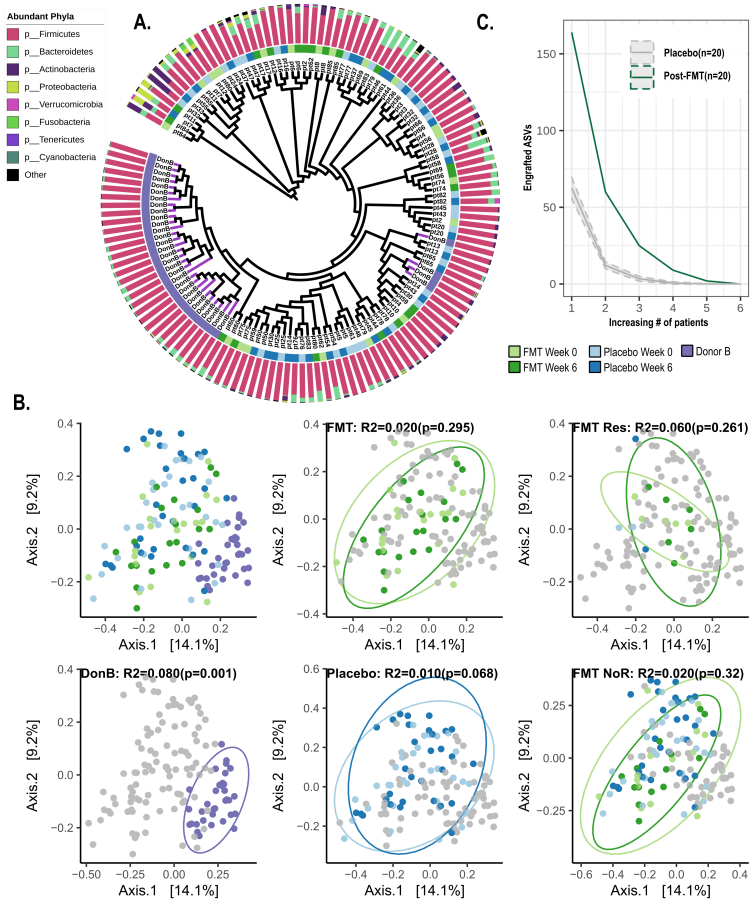

FIGURE 3.2: **16S rRNA gene sequencing does not indicate a common microbial shift following FMT**. The microbial composition of 51 patients who either randomly received FMT from donor B or placebo treatment using 16S rRNA gene amplicon sequencing. **A**. The UPGMA tree of Bray-Curtis distances for all the samples collected (patients and donor B). The light and dark green colours in the inner ring show baseline and week six samples, respectively, collected from patients who randomized to FMT treatment. The light and dark blue colours show baseline and week six samples, respectively, collected from patients who randomized to placebo treatment. The outer layer shows the taxonomic composition of samples at the phylum-level. **B**. The top left panel shows the PCoA of Bray-Curtis distances for all samples. The bottom left panel shows donor B samples compared to all the other samples collected from patients. The middle panels show the same PCoA space, comparing only the samples collected from before and after FMT or placebo treatments. The right panels compare samples collected prior and post-FMT in responder (top) and non-responder (bottom) patients. **C**. Donor B's ASVs that were commonly engrafted across an increasing number of patients post-FMT vs. placebo (as a control).

3.3.2 Shotgun metagenomics does not indicate consistent microbial engraftment across FMT responders

To increase the taxonomic resolution and to investigate the functional composition of the gut microbiota in the patients who received FMT from donor B, we conducted shotgun metagenomics on a subset of patients consisting of 6 responder and 4 non-responder FMT patients, 4 samples from donor B, and two samples from week 0 and week six from a non-responder patient who received placebo treatment (n=11 patients, 1 donor; 26 samples total, Fig. 3.1C). In order to assess whether there were microbial community-wide changes following FMT, we first identified the taxonomic composition of samples. Using Metaphlan2, we measured the Bray-Curtis and Aitchison beta-diversity distances between each sample and visualized the results via a PCoA (Fig. 3.3A) and UPGMA tree (Supplementary Figure A2.2A). Our results showed that there was not a significant community-wide change post-FMT either in all patients, or only those who responded to FMT (ANOVA non-parametric test based on Bray-Curtis distance; FMT R2 =0.02, p-value=0.9 Fig. 3.3A; Responder R2=0.04, p-value=0.9 Supplementary Figure A2.2C; or based on Aitchison distance; FMT R2=0.03, p-value=0.09; Responder R2=0.05,p-value=0.9). In addition to taxonomic composition, we also assessed community-wide changes based on identified microbial gene families following FMT. Visualising the Bray-Curtis and Aitchison metrics via a PCoA (Fig. 3.3B) and UPGMA tree (Fig. A2.2B) we did not observe a significant community-wide shift post- FMT (ANOVA non-parametric test based on Bray-Curtis distance; FMT R2 =0.03, p-value=0.9 Fig. 3.3B; Responder R2=0.04, p-value=0.8 Supplementary Figure A2.2D; or based on Aitchison distance; FMT R2=0.03, p-value=0.09; Responder R2=0.04,p-value=0.8).

Similar to the 16S rRNA gene sequencing results, we were not able to find a global microbial community shift post-FMT either at the strain or gene family level (Fig. 3.3A,B). To assess whether a group of microbial strains or gene families commonly engraft in patients who respond to FMT, we detected donor B's strain and gene families transferred

to patients following FMT using shotgun read-based metagenomics (see Methods). We identified 40 strains and > 600 gene families that were engrafted in ≥ 1 patients following FMT (Fig. 3.3C). Interestingly, there were 2 strains and 131 gene families that commonly engrafted in ≥ 3 patients post-FMT, and that were detected post-FMT in both responder and non-responder patients; none of them were present in the placebo treated patient (Supplementary Figure A2.2E and F). These shotgun metagenomic analyses show that FMT induces changes in the patient's microbiota strain and gene composition. However, these changes are not specific to those who went into remission following FMT.

We observed that the microbial composition of donor B using taxonomic and gene family composition is significantly different compared to all samples collected in this cohort (ANOVA non-parametric test based on taxonomy using Bray-Curtis distance; R2 =0.08, p-value=0.01 Fig. 3.3A; based on gene families R2=0.07 p-value=0.02 Fig. 3.3B). We concluded that engraftment occurs but is not specific to those who responded to FMT treatment. We need higher resolution data to investigate microbial engraftment following FMT and assess whether donor B could drive a signature of microbial changes across responder patients.

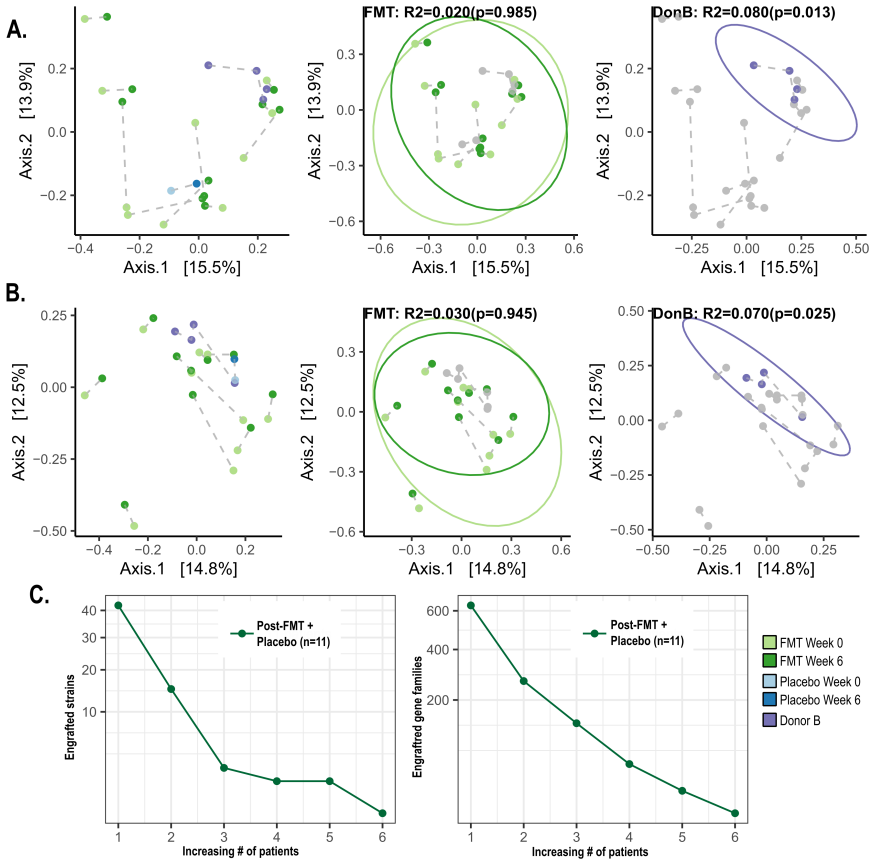

FIGURE 3.3: **Shotgun metagenomics shows microbial engraftment but is not specific to reponder patients**. The taxonomic and functional composition of 10 patients who received FMT from donor B and a patient on placebo treatment using shotgun metagenomics. **A**. The PCoA of Bray-Curtis distances based on the taxonomic composition of assigned reads. Dotted lines connect samples collected at week 0 and week 6 for each individual. **B**. The PcoA of Bray-Curtis distances based on the composition of known gene families in each sample. Donor B's strains (**C**.) and microbial gene families (**D**.) engrafted across increasing number of patients post-FMT.

3.3.3 Culture-enriched metagenomics improves the quality of *de novo* assembly and taxonomic binning

Detection of microbial engraftment may require greater microbial resolution (e.g., genomes and genes). It is often challenging to determine if the donor is responsible for the observed microbial changes post-FMT or if they instead represent patient's own microbiota shared with donor. *De novo* metagenomic approaches could provide a better resolution of the community, and it may be more suitable than amplicon or read-based metagenomic techniques for tracking strain or gene-level changes. However, the quality of genome assembly and taxonomic binning of *de novo* metagenomics can be poor (Sieber et al. 2018), and recovering the low-abundant taxa is challenging (Sczyrba et al. 2017). Previous studies have shown that the human microbiota is culturable (Lagier et al. 2012; Rettedal et al. 2014; Lau et al. 2016; Lagier et al. 2016; Lewis et al. 2021), and that a combination of *de novo* metagenomics and comprehensive culturing can result in increased observed microbial diversity (Forster et al. 2019; Whelan et al. 2020). Specifically, culture-enriched metagenomics could detect lower abundant organisms before FMT that are often missed by culture-independent methods, potentially improving detection of microbial engraftment following FMT. As such, we hypothesize that combining microbial *de novo* techniques with culture-enriched metagenomics of the donor B microbiota would improve microbial gene and strain recovery from this donor compared to the commonly used direct metagenomic approach (DMG). To test this hypothesis, a fresh fecal sample from donor B was plated on 33 media types under both aerobic and anaerobic conditions. 16S rRNA gene amplicon sequencing was conducted on each plate condition (66 culture conditions in total) as well as on the fecal sample itself (as described previously (Lau et al. 2016)). The minimum number of plate conditions with adequate ASV diversity to recapitulate the diversity of the donor B fecal sample were identified using the previously established plate coverage algorithm (PLCA, (Whelan et al. 2020)). Shotgun metagenomics was performed on 13 media conditions

as selected via the PLCA (Fig. 3.1D). We then compared the *de novo* assembly and microbial binning quality for a single fecal sample collected from donor B via DMG and CEMG approaches.

DMGs resulted in 35,577 assembled contigs (> 2.5 k) accounting for ∼340 Mbps of the assembled data, where as the same number of contigs in CEMG captured ∼620 Mbps (Supplementary Figure A2.3A, dashed lines). CEMG assemblies resulted in longer contigs compared to DMGs. Consequently, these longer contigs enhanced the *de novo* gene predictions and generated more (132 vs. 49) complete metagenome-assembled genomes (MAGs; > 70% completion and < 10% contamination) in CEMG compared to DMG (see Methods; Supplementary Figure A2.3A–C). To assess the CEMG approach, we mapped raw reads from both DMG and CEMG to the assembled MAGs from both approaches. CEMG recovered 83 more MAGs than DMG; however, most of these MAGs were present in the DMG based on the short-read mapping coverage, indicating that these results are not due to contamination but instead increased sequencing resolution (Supplementary Figure A2.3B). The increased number of MAGs in CEMG are not derived from a particular group of bacteria, but instead are enriched in all families in proportion to the original abundance in the DMG approach (Supplementary Figure A2.3B,C).

To examine the quality of assembled MAGs, we selected 40 homologous MAGs in DMG and CEMG based on their position in the phylogenetic tree and compared their genome size. We found that 24 of 40 (60%) of MAGs include more genetic information in CEMG compared to DMG (Supplementary Figure A2.3D). We concluded that CEMG enhanced the *de novo* assembly of genes and MAGs for intestinal microbiota and set to use this approach to detect and establish microbial engraftment following FMT.

3.3.4 High resolution mapping of the donor microbiota shows microbial genome engraftment following FMT

To refine the composition of donor B's microbiota regardless of temporal variations, we built a comprehensive database of donor B's microbiota using a co-assembly of four longitudinal DMG samples as well as one CEMG. In order to build a more comprehensive library of MAGs from donor B, we used the standard minimum 50% completion and maximum 10% contamination cutoffs Bowers et al. 2017 which identified 255 MAGs out of a total of 447 bins (Fig. 3.4A). To track donor B MAGs following FMT, we mapped raw shotgun reads from 11 patients (6 responders, 4 non-responders, and 1 placebo; 22 samples) to 255 donor B MAGs both before and after FMT. We then classified the donor B MAGs into *five microbial detection categories* by comparing the genomic coverage and single nucleotide variant (SNV) information from before and after FMT for each patient. These groups include: 1. **Unique to Donor** (Donor B MAGs that didn't transfer after FMT), 2. **Shared** (Donor B MAGs that are present in a patient both before and after FMT), 3. **Engrafted** (Donor B MAGs that were not detected before FMT but are present after FMT), 4. **Strain Replacement** (Donor B MAGs that replaced a patient's strain after FMT), 5. **Lost** (Donor B MAGs detected in a patient before but not following FMT).

"Unique to Donor" was the most abundant *category* across patients (204 / 255, on average). These MAGs were not present (defined as >60% of a MAG that has at least 1X coverage) in any patient sample (Fig. 3.4B, Supplementary Figure A2.4A). "Shared" and "Engrafted" were the second (31 / 255, on average) and third (10 / 255, on average) abundant categories, respectively. In order to assess if the MAGs present following FMT were a patient's own strains or if they were replaced by strains from the donor, we assessed the number of SNVs per kbp for "Shared" MAGs. Based on the number of SNVs identified, on average, 4 MAGs were categorized as "Strain replacement" (see

Methods, Fig. 3.4C, Supplementary Figure A2.4A) while most represented patient specific strains. We also identified a group of "Lost" MAGs strongly delineated in only one patient (pt79). Most MAGs (252 out of 255) were not present in patient 25 (microbial placebo treatment). These MAGs were classified as either "Unique to Donor" (n=170) or "Shared" (n=80) but we observed 2 "Strain replacement", a single "Engrafted", and a single "Lost" MAGs, highlighting the low margins of error of this model (Fig. 3.4B, Supplementary Figure A2.4A).

Using the microbial transfer model, we showed that most donor B MAGs were not transferred or present in patients after FMT. Although patients 4 and 10 (remission following FMT) illustrated the highest number of engrafted and replaced MAGs, we did not find an overall difference between responders and non-responders in terms of the total number of donor B MAGs transferred after FMT (Fig. 3.4B). Engrafted / replaced MAGs belong to 5 different bacterial phyla (as well as a single MAG assigned to Euryarchaeota) and the most abundant families in these MAGs were Lachnospiraceae, Osillospiraceae, Ruminococcaceae, Bacteroidaceae, and Acutalibacteraceae. Interestingly, we were able to detect 103 and 9 MAGs from donor B that engrafted / replaced in $\geq$ 1 and $\geq$ 3 patients, respectively, following FMT (Fig. 3.4D). However, these MAGs were not specific to responder patients (except a single MAG, M300, that we were only able to assign to Lachnospiraceae family; Fig. 3.4E). We concluded that the microbial shift at the genomic level is patient-specific; and that there are some specific strains from donor B which were able to engraft in $\geq$ 3 patients (n=9, Fig. 3.4D, E), regardless of whether they responded to FMT or not.

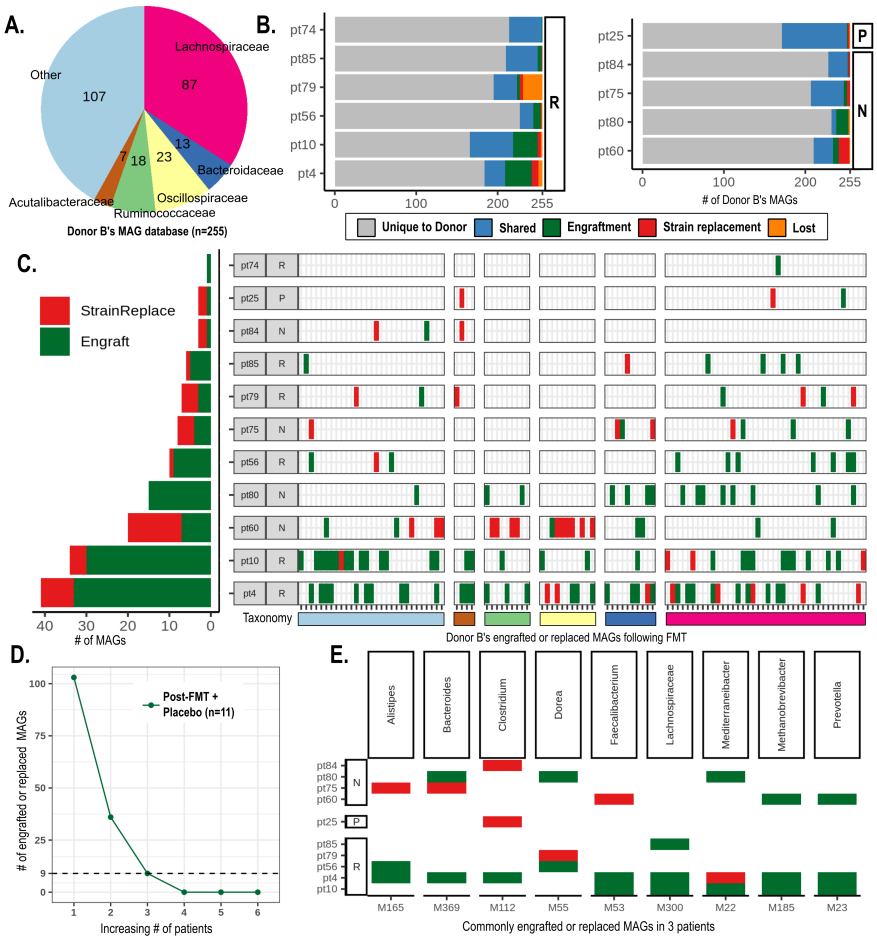

FIGURE 3.4: **High resolution genome-resolved metagenomics shows microbial genome engraftment and replacement following FMT**. **A.** *De novo* assembled donor B MAGs (n=255) via culture dependent and independent metagenomics, classified to the Family-level. **B.** Donor B's MAGs classified into different categories post-FMT in each patient using genomic coverage and single nucleotide variability. **C.** Comparison of the engraftment and strain replacement events across all patients post-FMT. The list of patients is sorted from the lowest to the highest number of engrafted and replaced MAGs. The taxonomy of each MAG is shown at the genus and phylum levels. **D.** The number of engrafted MAGs across an increasing numbers of patients. **E.** 9 donor B's MAGs commonly engrafted or replaced in ≥3 patients post-FMT.

3.3.5 A signature of gene engraftment in patients who responded to FMT

The CEMG approach allowed us to refine 255 MAGs from a single FMT donor, but tracking these MAGs alone does not provide a comprehensive assessment of the microbial dynamics after FMT. In highly related strains (or species) the presence or absence of only a few genes can correspond to a divergent functional profile (Karcher et al. 2021; Rousset et al. 2021). Subsequently, the detection of these strain replacement events following FMT becomes more challenging and required in-depth assessment of bacterial genes in addition to genomic coverage and SNV variability.

We were interested in identifying genes associated with engraftment and/or response to FMT. Because of the inherit challenges of metagenomic binning, we focused on the microbial genes assembled from donor B independent of their MAG/bin assignment, assembling and predicting 1,130,000 genes. Using stringent coverage thresholds (see Methods), we detected 139,535 (12%) genes that were potentially engrafted in $\geq$ 1 patient. In this model, engrafted genes are defined as those whose coverage before FMT were 0 but were $\geq$ 5X covered after FMT (see Methods). While many of the engrafted genes identified vary across patients, we identified a set of genes that commonly engraft across multiple patients. When we compared all 139,535 engrafted genes across all patients, 13,092 (9%), 267 (0.2%), and 7 (0.005%) genes commonly engrafted for at least 2, 3, and 4 patients respectively (Fig. 3.5A).

Interestingly, 265 of 267 ($\geq$ 3 patients) and all 7 ($\geq$ 4 patients) genes which commonly engrafted were specific to patients who responded to FMT (Fig. 3.5A, B). In contrast, only two of these genes were found in non-responder patients and none engrafted with placebo treatment (Fig. 3.5B). 43% of these genes belong to Lachnospiraceae (19% *Dorea*, 14% *Blautia*, 10% Other), and 21% Ruminococcaceae (11% *Faecalibacterium*, 10% Other); we were unable to predict the taxonomic origin of 30% of these genes. From

the 267 commonly engrafted genes, 50% are found in MAGs, 12% in bins and, 38% were not assigned a bin/MAG. Most of these genes (on average ∼500bp) are not well characterized (51% hypothethical proteins; (Fig. 3.5C). The top categories of Clusters of Orthologous Groups (COGs) predicted for these genes are transcription, translation, amino acid transport metabolism, carbohydrate transport metabolism, and coenzyme transport metabolism (Fig. 3.5C). 34% (n=93) of these genes have a corresponding protein in the Pfam database (Supplementary Table A2.1) The top predicted molecular functions across these genes are DNA binding (e.g., Cold-shock, antitoxins, Cro/C1-type HTH, Sigma-70, and Transposase IS200), ATP binding and ATPase activity (e.g., Type II/IV secretion system, Histidine kinase-like, Phosphomethylpyrimidine, AAA domain, ABC transporter), and hydrolase, kinase, peptidase activity (Supplementary Table A2.1). Some of these genes are part of mobile genetic elements (MGEs) or lysogenic phages that may improve the host strain's ability to compete in the gut environment (Rodriguez-Beltran et al. 2020; Koonin et al. 2020). For example, ATPase domains associated with relaxases, and ATP binding cassettes associated with DNA mobilization complexes are suggestive of mobile genetic elements (MGEs) (Coyne et al. 2014). The engraftment of these genes only in responder patients implicates the potential significance of these genes in the response to FMT treatment.

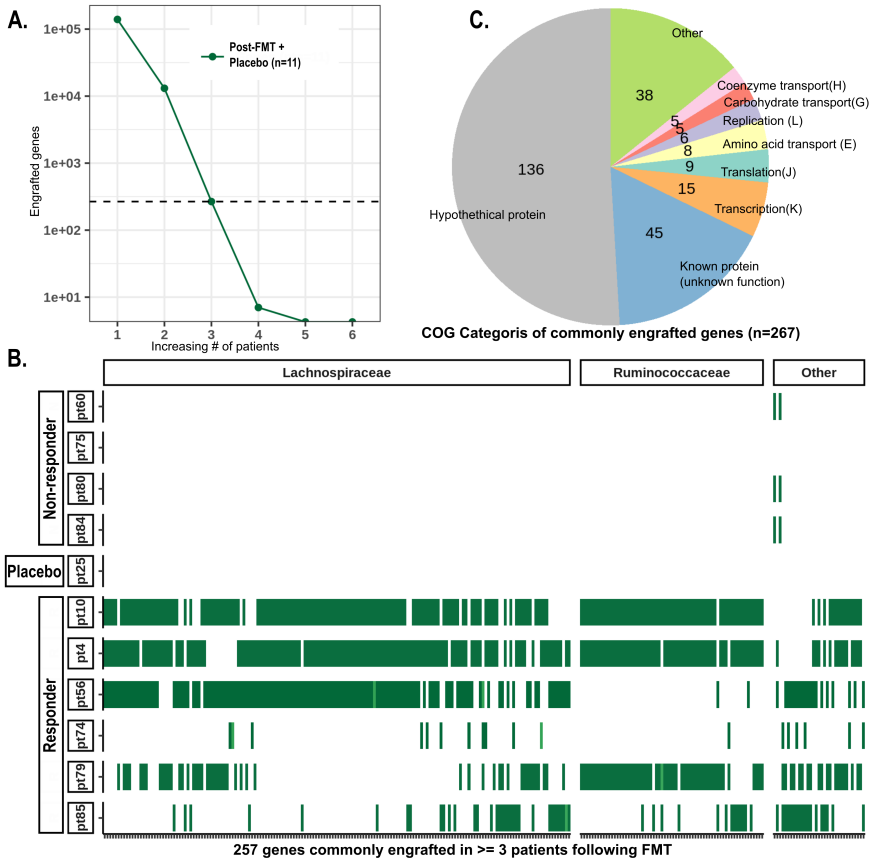

FIGURE 3.5: **Patients who responded to FMT show a signature of microbial gene engraftment.** A comparison of commonly engrafted genes across patients post-FMT. **A**. The number of commonly engrafted genes across an increasing numbers of patients. **B.** 267 genes are commonly engrafted in $\geq$3 patients. The taxonomic composition of these genes are shown at the Family- and Genus-level. A gene was called engrafted if it was not covered before FMT but was detected $\geq$5X coverage after FMT. **C.** The functional composition of commonly engrafted genes using categories of clusters of orthologous groups (COG) database.

3.3.6 The commonly engrafted genes identified in responder patients are strain specific

Of the 267 commonly engrafted genes (in $\geq$ 3 patients) 115 were associated with 3 genera (*Blautia*, *Dorea*, and *Faecalibacterium*). In order to test if these genes were strain-specific, we examined 402 whole-genome sequences (WGS) of *Blautia sp.*, *Dorea sp.*, and *Faecalibacterium prausnitzii* strains. This included 306 WGS from NCBI RefSeq (65 *Dorea*, 98 *F.prausnitzii*, and 143 *Blautia*) and 96 isolates from our lab strain collection (30 *Dorea*, 1 *F.prausnitzii*, and 67 *Blautia*). We constructed phylogenetic trees based on ribosomal proteins (independent of the commonly engrafted genes) and mapped the commonly engrafted genes to the phylogeny (Supplementary Figures A2.5, A2.6).

49% (47 of 95) of the genomes for *Dorea* were assigned to *D. longicatena*, and 39% (37 of 95) to *D. formicigenerans*; 11% of these genomes are not taxonomically classified at the species-level (Supplementary Figure A2.5A). From the 47 *D. longicatena* genomes, the phylogeny shows two distinct clades (Fig. 3.6A). All 15 genomes of the *D. longicatena* B clade contains $\geq$ 50 *Dorea* specific commonly engrafted genes, indicating strain-specificity of commonly engrafted genes in this species. In order to better understand the function of these commonly engrafted genes in *D. longicatena*, we mapped metagenomic read information to Isolate 14 — a sequenced strain within our lab collection that lies within clade B and was selected as the *Dorea* representative strain of commonly engrafted genes — and used stringent coverage information to detect commonly engrafted genes (0% of 1x detection before FMT with 5X coverage following FMT across $\geq$ 3 patients). Our results show that 42 bacterial genes commonly engrafted in $\geq$ 3 patients after FMT. The commonly engrafted genes present in Isolate 14 are not present in a particular gene neighbourhood and are distributed across the genome. Particularly, we identified glycan biosynthesis; sucrose metabolism pathway among the list of these proteins in Isolate 14. This supports that a specific *Dorea* strain from donor B replaced the strains in patients 4, 10, and 56 strains after FMT.

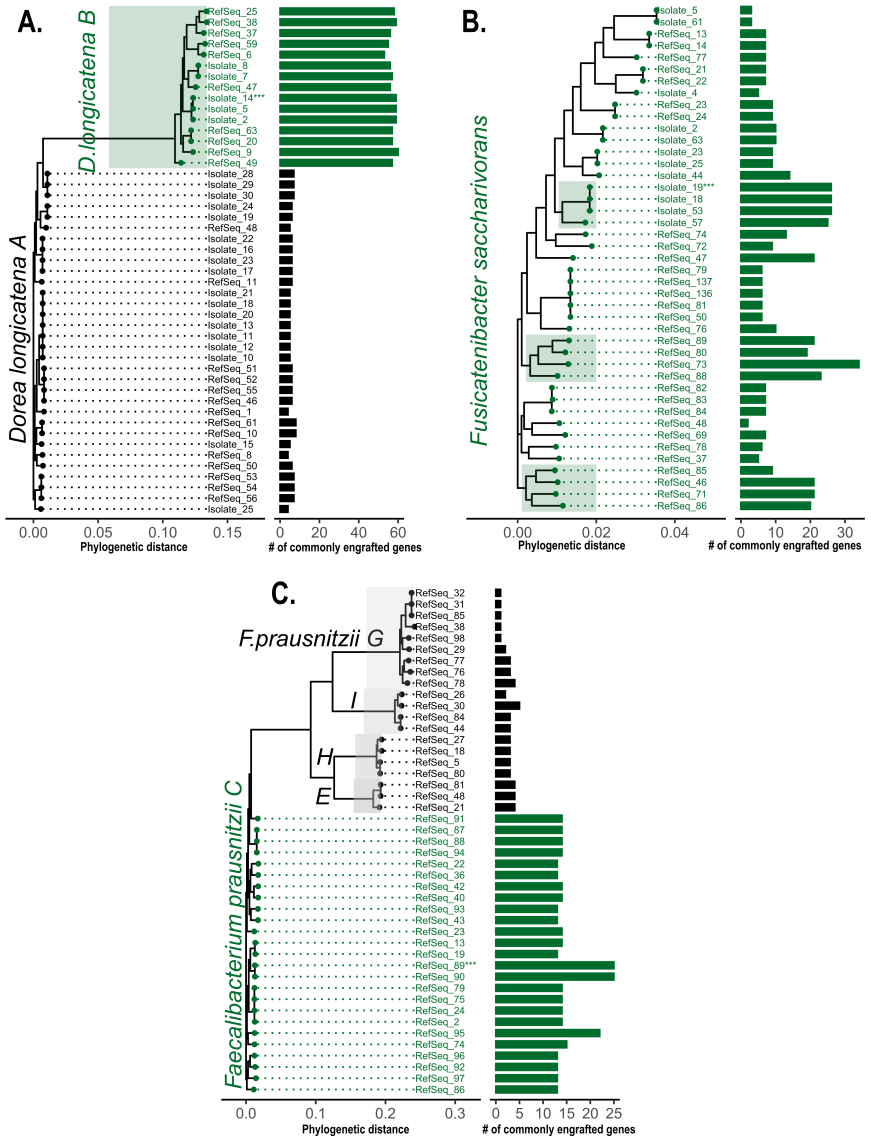

FIGURE 3.6: **The commonly engrafted genes are strain-specific**. A phylogeny of available strains in NCBI (**RefSeq #**) as well as Surrette lab whole genome collection (**Isolate #**) constructed per species. The number of commonly engrafted genes identified in each genome are shown in **A.** *Dorea longicatena* (n=47), **B.** *Faecalibacterium prausnitzii* (n=45), and **C.** *Fusicatenubacter saccharivorans* (n=43) phylogenies. A representative strain of commonly engrafted genes (rsCEGs, annotated with $***$) selected for each phylogenetic tree.

The phylogenetic tree of all 210 *Blautia* genomes (Supplementary Figure A2.6) revealed a single lineage that contained the engrafted genes. The taxonomy of this group

was unclassified in NCBI RefSeq but assigned to *Fusicatenibacter saccharivorans* in the GTDB taxonomy database (Parks et al. 2020). Interestingly, we identified 11 genomes with $\geq$ 15 commonly engrafted genes that all belonged to a particular phylogenetic clade (Fig. 3.6C). The identified commonly engrafted genes are distributed throughout this species but three groups are enriched in these genes and suggest that the commonly engrafted genes associated with *F. saccharivorans* represent engraftment of a specific donor B strain in FMT patients. The genomic coverage of Isolate 19 — a sequenced strain within our lab collection selected as the *F. saccharivorans* representative strain of commonly engrafted genes — across donor B samples showed that this particular strain is stable and consistently present in donor B (2012–2017). However, the genomic coverage of Isolate 19 in UC patients is variable, which implies strain replacement following FMT. To assess strain replacement of the patient's strain after FMT, we have applied our engraftment model to identify the commonly engrafted genes following FMT. Our results show that 135 genes were engrafted in $\geq$ 3 patients. Note the 135 engrafted genes predicted using the genome of Isolate 19 is greater than the 38 genes identified from the Donor B CEMG database. This reflects the stringency of data included in the database limiting it to contigs >2.5kb. The majority of these genes are present in a 40 kbp gene cluster in patients 4, 56, and 85. These bacterial genes are part of various pathways such as phospholipid metabolism, L-tryptophan and L-histidine biosynthesis. In order to determine whether these genes are the result of strain replacement or whether they were transferred to the patients microbiota via horizontal gene transfer (HGT), we selected adjacent flanking regions surrounding this gene cluster (20,000bp on each side) and compared SNVs across patients. High SNV frequency in flanking regions suggest the gene cluster originates from HGT while little/no SNV frequency suggest strain replacement (Supplementary Figure A2.7). The comparison of genomic coverage and SNV variability across donor and patients samples does not provide any evidence for bacterial HGT (Supplementary Figure A2.7). In contrast, we observed increased

variability for this region in patient 25 (placebo treatment) at week six and low coverage in patient 75 (non-responder) following FMT. The data presented here is consistent with *F. saccharivorans* strain replacement after FMT, specifically in 3 of the responders (Supplementary Figure A2.7).

Faecalibacterium prausnitzii genomes are highly diverse and this was reflected in the 99 strains we analysed (Supplementary Figure A2.5B). The ribosomal protein based phylogenetic tree resolved into 15 *F. prausnitzii* clades. Recently, 22 *Faecalibacterium*-like clades were refined using a larger dataset and metagenomic binning approach (De Filippis et al. 2020). We found that 25 of the 99 genomes (part of a single clade) contained ≥ 15 commonly engrafted genes from donor B (Fig. 3.6B). Similar to the *Dorea* and *F. saccharivorans* collections, we showed that the donor B commonly engrafted genes are strain-specific among a collection of *F. prausnitzii* genomes. The majority of the commonly engrafted genes in RefSeq 89 — a selected representative strain of commonly engrafted genes — are predicted to be lysogenic phage genes corresponding to uncharacterised proteins, most located in a single gene-cluster 60 kbps in length in RefSeq 89. To assess if the identified lysogenic phage was transferred via HGT, we used SNV variability information similar to the previously described method. Two flanking regions adjacent to the 60 kbps commonly engrafted region were selected and filtered for only donor B conserved base positions. Given the identity of reported SNVs from before and after FMT in the extracted flanking region, we argue that the commonly engrafted genes identified in this genome are the result of bacterial HGT and not strain replacement (Supplementary Figure A2.8). The comparison of identified SNVs compared to the donor B samples showed patient-specific patterns of SNVs and gene coverage for all patients (Supplementary Figure A2.8). The lysogenic phages are possibly a source of selective pressure for high strain diversity in *Faecalibacterium prausnitzii*-like strains (Cornuault et al. 2018). The identified lysogenic phage in RefSeq 89 appears to be strain-specific and possibly provides a unique advantage to their bacterial host to succeed in

bacterial competitions after FMT.

3.3.7 The 3 donor B strains identified in FMT responders are depleted in IBD patients

Our results showed that the microbial strain replacement following FMT can be seen for a group of accessory genes within a genome. To explore whether the deficit of these genes is associated with disease activity in a strain- or species-specific manner, we developed strain- and species-specific markers for the three representative strains of commonly engrafted genes. We used commonly engrafted genes as strain-specific markers for each representative strain and developed species-specific core-gene markers of similar size (50 kb gene clusters) using pangenome analysis (Supplementary Figure A2.9, see Methods).

We first validated the accuracy of these markers with our current data (Supplementary Figure A2.10, A2.11). This approach allows us to separately determine the presence of the species and the specific set of engrafted genes that define specific strains of interest. The detection of conserved markers (1X coverage of $\geq$ 80% gene clusters) alone shows the presence at the species level. However, the detection of both conserved and commonly engrafted gene markers in a metagenomic sample indicates the presence of the representative strain containing the commonly engrafted genes. For example, we verified that all donor B samples contain only the strains of interest and strain-specificity increased in UC patients post-FMT (30% in *D. longicatena*, 30% in *F. prausnitzii*, and 50% in *F. saccharivorans*; A2.11).

A limitation of our study was the small number of samples examined. In order to explore the relevance of these commonly engrafted genes more broadly, we examined the distribution of these species- and strain-specific genes in IBD patients compared to healthy controls from publicly available metagenomic data. We mapped reads from a metagenomic dataset (NCBI SRA ID: PRJNA400072) containing patients with UC

(n=76), Crohn's disease (CD, n=88), and healthy controls (n=56) (Franzosa et al. 2019) to the species and strain-specific markers (Fig. 3.7). For *D. longicatena* a decrease in prevalence with the disease was observed with the species-specific markers being detected in 82% of healthy controls and only 46% and 29% of UC and CD patients (Fig. 3.7B). A similar decrease in prevalence was observed for *F. prausnitzii* (present in 91% HC vs 70% and 25% UC and CD patients respectively) and *F. saccharivorans* (present in 91% HC vs 45% and 52% UC and CD patients respectively). Both strains with and without the commonly engrafted genes were depleted in the IBD patients (Fig. 3.7B). Interestingly, the strain-specificity decreased in UC patients 2.7 and 3.4 fold in *F. prausnitzii* and *F. saccharivorans*, respectively, compared to healthy individuals, while species-specificity only reduced 1.2 and 1.5 folds respectively for the same species.

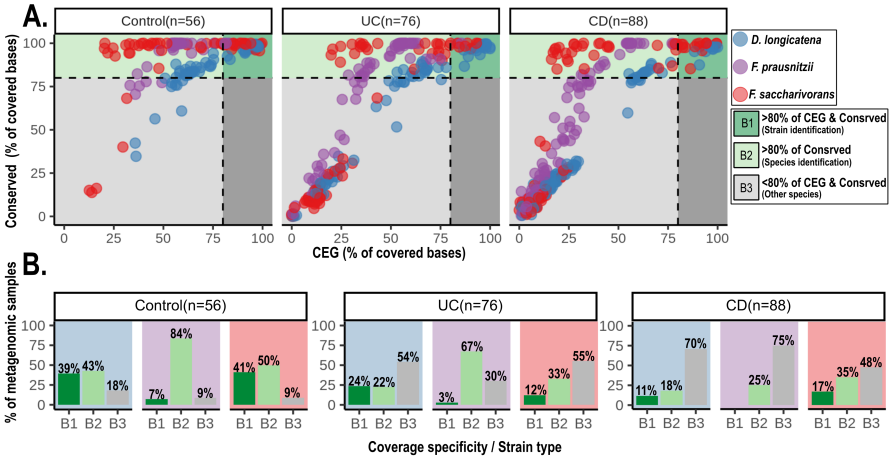

FIGURE 3.7: Tracking the representative strains of commonly engrafted genes in metagenomic samples using strain and species-specific markers. The specificity of *D. longicatena*, *F. prausnitzii*, and *F. saccharivorans* representative strains compared across metagenomic samples a publicly available metagenomic dataset (bottom row in each figure; healthy controls (n=56), UC (n=76), and CD (n=88)). **A.** Comparison of a conserved (species-specific) vs. commonly engrafted gene (strain-specific) cluster for each strain within each metagenomic sample. Each dot represents one genome in a metagenomic sample. **B.** The classified genomes from a metagenomic sample based on conserved and commonly engrafted gene's coverage percentage. Genomes with CEG=commonly engrafted gene and conserved gene cluster coverage $\geq$ 80% (dark green) and those with conserved coverage $\geq$ 80% (light green) in a metagenomic sample are labelled as B1 (strain-specific) and B2 (species-specific) respectively. The genomes with conserved region coverage $<$ 80% are labelled as B3 (other species). rsCEGs= representative strain of commonly engrafted genes from Figure 3.6

3.4 Discussion

FMT has recently gained attention as a treatment for patients with UC. The efficacy of this approach has been shown in the RCTs comparing FMT to placebo (Narula et al. 2017). FMT is not risk-free (DeFilipp et al. 2019); however, it can lead to more targeted microbial therapies by better understanding why some patients respond to FMT treatment while others do not. We are yet to recognize the role of the donor's microbiota in the successfulness of FMT for UC patients. While some past trials combined multiple donor microbiomes (Costello et al. 2019), others have reported donor-dependent efficacy (Moayyedi et al. 2015; Wilson et al. 2021). Within this article, we set out to study a successful FMT donor in an RCT for patients with UC. We have focused specifically on whether engraftment of a donor microbiota was associated with remission post-FMT.

16S rRNA gene sequencing does not provide enough resolution to track microbial engraftment (Fig. 3.2). We found that there was no signature of engraftment associated with response in patients receiving FMT. Moreover, it demonstrates the degree of noise in these analyses, as a significant number of "donor specific ASVs" were absent in patients at baseline but do appear in the placebo patients at the end of the study period. This is because there are always taxa present but below the level of detection that can confound the analysis. We have previously demonstrated this through extensive culture approaches where more taxa were detected by culture than by direct 16S rRNA gene sequencing (Lau et al. 2016).

Our read-based metagenomic analysis has shown that microbial engraftment was not specific to those who responded to FMT (Fig. 3.3). Similar to the previous studies using a marker-based metagenomic pipeline (Franzosa et al. 2018) to uncover microbial engraftment in UC Paramsothy et al. 2019; Chu et al. 2021, rCDI ((Smillie et al. 2018)), and obesity Wilson et al. 2021, we observed evidence of microbial engraftment post-FMT. However, there was not any association between engraftment and remission post-FMT.

These profiles that metagenomic markers and uniref90 gene-families have characterized do not represent unique donor strains or genes. In addition, approximately 40% of the metagenomic reads were not mapped to any gene-family marker in this approach (Supplementary Figure A2.2B). Many donor strains are closely related, and the presence/absence of only a few gene clusters in the accessory genome can distinguish these strains from each other.

De novo assembly of short metagenomic reads into contigs and MAGs provides a more robust resolution of the gut microbiota, and the effectiveness of this approach to track microbial engraftment has been shown previously (Lee et al. 2017; Watson et al. 2021). When we applied the same method to our culture-enriched collection of a donor B sample, we found that CEMG improved the quality of the *de novo* gene and genome recovery of gut microbiota compared to direct shotgun metagenomics (Supplementary Figure A2.3A). Culture-independent sequencing methods revolutionized our understanding of human intestinal microbiota (Dominguez-Bello et al. 2011), but they often miss low-abundant bacteria. For example, it was shown that culture-dependent 16S rRNA amplicon gene profiling recovers a greater number of OTUs compared to culture-independent approaches with the same depth of sequencing (Whelan et al. 2020). These low abundant bacteria could be essential in FMT treatments.

We observed that most donor B MAGs were not engrafted or replaced post-FMT, and those that transplanted showed a patient-specific pattern (Fig. 3.4). The engrafted or replaced MAGs indicated inconsistent engraftment across responder patients (Supplementary Figure A2.4). Further, we could only assign 52% of donor B's assembled base pairs into MAGs, even with the CEMG approach. Although metagenomic *de novo* assembly and taxonomic binning have seen recent algorithmic improvements, it's still a challenge to refine highly related strains from a complex microbial community (Yue et al. 2020). We argued that detecting strain engraftment and replacement events following

FMT requires an in-depth assessment of bacterial genes.

Tracking donor B's microbial genes identified 265 genes that are commonly engrafted in $\geq$ 3 responder patients (Fig. 3.5). 115 genes commonly engrafted post-FMT were associated with *Fusicatenubacter*, *Dorea*, and *Faecalibacterium*. These genes are species and strain-specific (Fig. 3.6, Supplementary Figure A2.5, Supplementary Figure A2.6). We identified a particular phylogenetic clade with the greatest number of commonly engrafted genes within each species. Our results showed that these engrafted genes were the result of strain replacement in *D. longicatena* and *F. saccharivorans* and that some of them share homology with MGEs. The commonly engrafted genes identified from *F. prausnitzii* likely represent a horizontal gene transfer event, and these genes were predicted to be within a lysogenic phage. MGEs and lysogenic phage are widespread among commensal intestinal bacteria. For example, recently, it implicated that strains secreting Bacterial ADP-ribosyltransferases (ADPRTs) associated with phage elements can positively select strain colonization than other closely related strains (Brown et al. 2021).

Our dataset was limited to samples prior- and post-FMT; however, recently, it was shown that the engraftment of a donor's strains post-FMT is stable for an extended period (Aggarwala et al. 2021). To assess the presence of these engrafted genes within the three strains in a larger cohort of IBD patients and healthy controls, we developed species and strain-specific markers for each representative strain and evaluated their accuracy using data from this study. We observed that the strains from donor B that replaced in FMT recipients in $\geq$ 3 patients were also depleted in IBD patients compared to healthy controls (Fig. 3.7). Similar to previously published data (Franzosa et al. 2019), we observed that *D. longicatena* and *F. prausnitzii* are depleted in IBD patients, however; distinguishing closely related strains belonging to *F. prausnitzii* and *F. saccharivorans* were crucial to assessing the relevance of these bacteria to disease activity. This suggests

that these genes are depleted in patients and implicates strains carrying these genes in successful FMT.

Our data is consistent with the engrafted strains having a fitness advantage over closely related strains of the same species. This advantage is associated with clinical response and implicate these strains in promoting remission. However, the fitness advantage may be manifested in a more healthy gut environment in which case this association with response may be a consequence and not a cause of remission. This study highlights the challenges in studying engraftment in FMT and the importance of strain level characterization. Using high resolution metagenomic data generated from culture-enriched metagenomics of the donor microbiome improves our ability to detect engraftment and demonstrates that large scale engraftment of donor microbes to patients is not occurring during FMT. Only a few engrafted strains are specifically associated with response across multiple patients and these strains may have therapeutic potential for designed microbiota consortia for FMT in ulcerative colitis.

Chapter 4

Longitudinal dynamics and transferability of crAssphage

4.1 Introduction

The human intestinal microbiota that contains bacteria, viruses, archaea, and fungi is highly linked to health and disease. Bacteriophages — bacterial viruses — predominated the human gut virome (Sutton and Hill 2019; Manrique et al. 2016). Despite this abundance, until recently, the human gut bacteriophages (phages) have been poorly characterized in relation to the rest of the human microbiome (Minot et al. 2013; Roux et al. 2015; Shkoporov et al. 2019; Shkoporov et al. 2022). Since bacteriophages infect only bacteria, they can alter the human gut microbiome through various implicated mechanisms, such as horizontal gene transfer (Chen et al. 2018) and elimination of their host (Avrani et al. 2012). Therefore, intestinal bacteriophages have an impact on human health (Norman et al. 2015).

crAssphage is the most abundant bacteriophage in the human gut, initially identified by metagenomics, and it is estimated that crAssphage is present in ∼40% of humans (Dutilh et al. 2014). This phage has a ∼95-97 kb circular, double-stranded DNA genome. crAssphage sequences are found in human fecal metagenomes in diverse populations globally and can be highly abundant. Recent studies have shown that crAssphage is one member of a wide range of crAss-like phages (Alpha, Beta, Gamma, Delta subfamilies, and 10 clusters) that exist in the human microbiome (Guerin et al. 2018). The crAssphage has since been found to be globally distributed, with strains reflecting the geographic distribution of human populations (Edwards et al. 2019). Sequencing of the crAssphage genome demonstrated that the phylogeography of crAssphage is locally clustered within countries, cities, and individuals (Edwards et al. 2019). Subsequently, crAssphage has been studied in a variety of environments, from infant gut samples, to patients with diarrhea, and in samples from healthy donors and fecal microbiota transplantation (FMT) recipients (Liang et al. 2016; Siranosian et al. 2020). Additional metagenomic evaluation has demonstrated that crAssphage is closely associated with

human fecal waste, and crAssphage has been used for human fecal source identification (Stachler et al. 2017; Karkman et al. 2019; Wu et al. 2020).

FMT involves the transfer of fecal matter from a healthy donor to a recipient in an attempt to restore microbiota diversity and composition. Currently, FMT is mostly used for the treatment of recurrent- *Clostridioides difficile* infection (rCDI), where it has been found to be highly effective. Studies have shown evidence for engraftment of donor bacteria into recipients (Smillie et al. 2018; Paramsothy et al. 2019; Wilson et al. 2021), but information about viral alterations after FMT treatment is limited (Lam et al. 2022). Sterile filtrates from donor stool, rather than fecal microbiota, can be sufficient to restore normal stool habits and eliminate symptoms after *Clostridioides difficile* (*C. difficile*) infections. Therefore, it is possible that bacteriophages are mediating some of the effects of FMT. FMT studies provide an opportunity to look at crAssphage engraftment and potentially at strain competition.

In this study, I used donor samples from an FMT randomized controlled trial (RCT) for patients with ulcerative colitis (UC) (Moayyedi et al. 2015) to examine long term crAssphage dynamics in a healthy donor (>5 years). Recipient samples from this FMT study and publicly available data (Draper et al. 2018) were used to study short-term dynamics and potential engraftment of donor crAssphage. Shotgun metagenomics and assembly was used to identify crAssphage in each sample and read mapping was used to characterize population heterogeneity and crAssphage transfer from donor to recipients.

4.2 Methods

4.2.1 Study design and sample collection

Five longitudinal stool samples were collected from a single healthy individual (donor B; 2012−2017, see Table 4.1). This individual was an FMT donor for a RCT of FMT for UC patients (Moayyedi et al. 2015). We have conducted shotgun metagenomics on paired

samples (pre-and post-FMT) from 10 patients who received FMT treatment from donor B, and a single patient who received placebo treatment (11 x 2= 22 samples in total). We also collected cross-sectional stool samples from seven healthy donors (SHCM1−6 and SHCM15). Publicly available viral metagenomic dataset (PRJNA446038) from sequence read archive (SRA) was also investigated. This dataset contains viral metagenomic data from donors and rCDI patients who received FMT from healthy donors with follow up samples up to 12 months (Draper et al. 2018).

4.2.2 DNA extraction and metagenomic library preparation

Briefly, 0.2 g of fecal matter was mechanically homogenized using 2.8mm ceramic beads, 0.1mm glass beads in 800 μL of 200 mM NaPO 4 (pH 8) and 100 μL of guanidine thiocyanate-EDTA-N-lauroyl sarcosine. This was followed by enzymatic lysis of the supernatant using 50 μL of 100 mg/mL lysozyme and 10 μL of 10 mg/mL RNase A for one hour at 37° Then, 25 μL of 25% sodium dodecyl sulfate (SDS), 25 μL of 20 mg/mL proteinase K, and 75 μL of 5 M NaCl was added, and incubated for one hour at 65°C. DNA was purified using the MagMAX Express Magnetic Particle Processor (Thermofisher, Burlington, ON) as per manufacturers instructions. DNA was standardized to 5 ng/ μL and sonicated to 500 bp. Using the NEBNext Multiplex Oligos for Illumina kit (New England Biolabs), DNA ends were blunted, adapter ligated, PCR amplified, and cleaned as per manufacturers instructions. Library preparations were sent to the McMaster Genome Facility, and sequenced using the Illumina HiSeq platform.

4.2.3 *De novo* assembly of crAssphage genomes from metagenomics

Low-quality reads and sequencing primers were removed using Trimmomatic (Bolger et al. 2014). Samples were assembled from paired-end reads using metaSPAdes (Bankevich et al. 2012), except for one sample (donor B 2012) that was assembled via SPAdes

(Bankevich et al. 2012) using single-end reads. A local BLAST database for each assembled library was generated and searched for sequences with a minimum 90% pident against the uncultured crAssphage reference in NCBI (NC_024711.1), which belongs to the alpha subfamily from cluster one of crAss-like phages (Guerin et al. 2018). These hits were aligned against this reference genome via a circular genome aligner (CSA; Fernandes et al. 2009), and they were reverse completed in case they were in the opposite strand. Based on their alignment to the reference genomes, the contigs with the correct orientation were merged as a draft crAssphage genome. Gene prediction and annotation for each phage genome was carried with Prokka (Seemann 2014). The reference crAssphage (NC_024711.1) was used for annotation for consistency with previous studies.

4.2.4 Assessing crAssphage variability in metagenomic samples

Trimmed shotgun reads from each sample were mapped using bwa-mem (Li and Durbin 2009) to: 1) the reference crAssphage (NC_024711.1) genome, 2) *de novo* assembled crAssphage genomes from each sample, and 3) the crAssphage genome assembled from donor B's from 2013 sample. Then I used samtools (Li et al. 2009) to get coverage and breseq (Deatherage and Barrick 2014) to identify SNPs. I merged the coverage and SNP information for every single base position in R v. 4.0.3. Figures containing gene annotation and genome coverage were generated using tidyverse (Wickham et al. 2019) and gggenomes packages. A phylogenetic tree of *de novo* assembled crAssphage genomes was generated by whole-genome alignment using mafft (Katoh et al. 2002) and approximately-maximum-likelihood model via fasttree (Price et al. 2010), and visualized using the gggenomes package in R v. 4.0.3.

4.2.5 crAssphage host in donor B samples

Metagenomic reads from all donor B samples were mapped to the reference crAssphage using bwa-mem (Li and Durbin 2009). Total and mapped read numbers were parsed from samtool's flagstat (Li et al. 2009) outputs, and the relative abundance of crAssphage was calculated for each sample. MetaPhlAn3 (Nousias and Montesanto 2021) was used to profile the relative abundance of microbial species for all donor B samples. I used a Spearman rank-sum test to estimate the association between crAssphage and bacterial species for each sample in R v. 4.0.3. All figures were made in R v. 4.0.3 using tidyverse package.

4.2.6 crAsSNPer pipeline for accurate detection of crAssphage engraftment

Using the public dataset (PRJNA446038), I assembled donor D3's crAssphage genome by co-assembly of samples F0 and F1 using SPAde (Bankevich et al. 2012). I then used the crAsSNPer pipeline to detect crAssphage engraftment in Draper et al. 2018 dataset and in our own FMT dataset, as follows.

The metagenomic reads were mapped to the appropriate (donor B for our data, donor D3 for the downloaded data) *de novo* assembled crAssphage genome using bwa-mem (Li and Durbin 2009). Samtools (Li et al. 2009) was used to calculate genome coverage and mean depth of coverage for each sample. Samples with coverage over $\geq 90\%$ of the crAssphage genome's length and mean coverage depth ≥ 10X were selected as crAssphage positive, and the lowest mean coverage depth was identified across these samples. Reads mapping to the crAssphage genome were extracted from the bam file and converted to a fastq file for each sample using samtools (Li et al. 2009). These reads were subsampled with replacement to the lowest coverage depth using samtools -s (Li et al. 2009) 20 times for each sample. breseq (Deatherage and Barrick 2014) was used to identify SNPs across these samples compared to their appropriate assembled genome (donor B or D3). breseq

output tables were merged together using breseq gdtools ANNOTATE (Deatherage and Barrick 2014). The final SNP table was merged with the metadata from the study and figures were made in R v. 4.0.3 using the tidyverse (Wickham et al. 2019) package. The median of the resampling for each sample was used to model the among-individual differences in SNP counts for the downloaded data. Resampled data from our FMT dataset were plotted directly. The tukey method was used for the pairwise comparisons to identify significantly different estimates.

TABLE 4.1: Metagenomic samples from healthy donors and ulcerative colitis subjects examined for crAssphage.

Number	Sample	Individual	Time	Status	p-crAssphage	Total reads	crAssphage reads	Assembly length	PCR result
1	B2012	donorB	2012	Healthy	positive	26146370	19477	96034	positive
2	B2013	donorB	2013	Healthy	positive	39780181	6692	96198	positive
3	B2016	donorB	2016	Healthy	positive	13875813	13440	96717	positive
4	B2017A	donorB	May 2017	Healthy	positive	61859261	103477	97496	positive
5	B2017B	donorB	Oct 2017	Healthy	negative	48934277	3	0	negative
6	SHCM1	SHCM1	single	Healthy	negative	37786351	997	0	
7	SHCM2	SHCM2	single	Healthy	positive	58502658	146075	93266	
8	SHCM3	SHCM3	single	Healthy	negative	86230076	3000	0	
9	SHCM4	SHCM4	single	Healthy	positive	57676099	97220	93982	
10	SHCM5	SHCM5	single	Healthy	negative	57548384	4	0	
11	SHCM6	SHCM6	single	Healthy	negative	80652914	0	0	
12	SHCM15	SHCM15	single	Healthy	negative	233990946	0	0	
13	PMCL380	pt4	pre-FMT	UC	negative	29879728	20	0	negative
14	PMCL385	pt4	post-FMT	UC	negative	25579014	11	0	negative
15	PMCL356	pt10	pre-FMT	UC	negative	22968602	2	0	negative
16	PMCL360	pt10	post-FMT	UC	negative	28442655	476	0	negative
17	PMCL656	pt25	pre-placebo	UC	negative	66881504	46	0	negative
18	PMCL657	pt25	post-placebo	UC	negative	85160380	8	0	negative
19	PMCL720	pt56	pre-FMT	UC	negative	12298708	0	0	negative
20	PMCL721	pt56	post-FMT	UC	negative	35411714	7	0	negative
21	PMCL726	pt60	pre-FMT	UC	negative	11991630	14	0	negative
22	PMCL727	pt60	post-FMT	UC	negative	14118620	36	0	negative
23	PMCL796	pt74	pre-FMT	UC	negative	14154374	60	0	negative
24	PMCL797	pt74	post-FMT	UC	negative	40102264	134	0	negative
25	PMCL800	pt75	pre-FMT	UC	negative	12548192	12	0	negative
26	PMCL801	pt75	post-FMT	UC	negative	11006750	13	0	negative
27	PMCL813	pt79	pre-FMT	UC	negative	12813300	0	0	negative
28	PMCL883	pt79	post-FMT	UC	negative	13720164	6	0	negative
29	PMCL817	pt80	pre-FMT	UC	negative	12136957	7	0	negative
30	PMCL818	pt80	post-FMT	UC	positive	14074341	10730	94148	positive
31	PMCL822	pt84	pre-FMT	UC	positive	11804105	297409	96661	positive
32	PMCL823	pt84	post-FMT	UC	positive	15044972	163677	96661	positive
33	PMCL824	pt85	pre-FMT	UC	negative	12252354	22	0	negative
34	PMCL825	pt85	post-FMT	UC	negative	14119902	15	0	negative

4.3 Results

Using a metagenomic dataset containing 34 fecal samples from healthy donors and UC patients (Table 4.1), I asked whether the crAssphage is variable in longitudinal samples

from a single healthy subject (donor B; 2012−2017; Table 4.1). I investigated crAssphage variability in terms of genome structure, abundance, and population (strain variability) using *de novo* assembly of the genomes, shotgun metagenomic reads counts, and single-nucleotide polymorphism (SNP) data, respectively. Since donor B provided fecal matter to a RCT of FMT for UC patients, I also asked whether donor B's crAssphage was engrafted in UC patients post-FMT. I used paired samples from UC patients (pt4−85; n=11; Table 4.1) and donor B to investigate these questions. I have focused on p-crAssphage — the original uncultured crAssphage deposited in NCBI (NC_024711.1, Dutilh et al. 2014) — which was classified as an alpha subfamily, cluster 1 of crAss-like phages (Guerin et al. 2018) and here I refer to it as reference crAssphage.

4.3.1 crAssphage dynamics in longitudinal samples from donor B

In order to test crAssphage dynamics, I investigated five longitudinal metagenomic samples from a healthy individual (donor B; 2012-2017). To do this, I mapped metagenomic reads from each sample to the reference crAssphage. I showed that the relative abundance of crAssphage has been variable in donor B over time (Fig. 4.1A). Interestingly, this phage was highly abundant (0.2% of metagenomic reads) in May 2017 (2017A) while it was completely absent in October of the same year (2017B;Fig. 4.1A). I found that the proportion of the crAssphage reference genome — covered 1X — increased from 93% (2012-2016) to 96% (2017A) and disappeared in 2017B sample (Fig. 4.1B) suggesting that the donor B's crAssphage was more similar to the reference genome in 2017A sample. I compared the PCR and metagenomic results for these five samples from donor B (Fig. 4.1C). These results showed that the designed primer sets are sensitive enough to detect the presence of crAssphage in a sample and we have used this to expand the number of donor B's samples used for tracking crAssphage.

4.3.2 crAssphage host bacteria in donor B

Definitive host bacteria for crAssphage have remained elusive although it has been demonstrated that crAssphage — can replicate in vitro in *Bacteroides intestinalis* (Shkoporov et al). From our metagenomic samples from donor B I looked for correlations in abundances in bacterial species and crAssphage over time. I was not able to associate the relative abundance of any of the *Bacteroides sp.* to crAssphage (Figs. 4.1D, 4.2A). However, I found that the donor B 2017B sample, which was negative for crAssphage, showed an increased relative abundance of a *B. vulgatus*. Instead, our results showed that *Eubaterium sp CAG 180* and *Roseburia intestinalis* were significantly associated (Spearman's correlation=1, p=0.02) with crAssphage (Fig. 4.2B) among all the bacterial species identified in these samples.

4.3.3 crAssphage variability in *de novo* assembled genomes

To compare the genomic structure of donor B's crAssphage over time, I have *de novo* assembled crAssphage genomes in samples collected from donor B. crAssphage contigs were identified in metagenomic assemblies and aligned to the reference genome to construct draft genomes. I found that the genomic organization of donor B's crAssphages was also variable. crAssphages assembled from samples taken 2012-2016 are most similar to each other while assembled crAssphage in 2017A is more similar to the reference crAssphage (Figs. 4.1E, 4.8). These differences are in the presence/absence of hypothetical proteins related to phage replication (putative dUTP), putative tail, and a single gene related to capsid (Fig. 4.1E). I assembled complete genomes in all crAss positive samples except donor B 2012 sample, which was a single-end sequencing and resulted in genomic re-arrangement (Fig. 4.8).

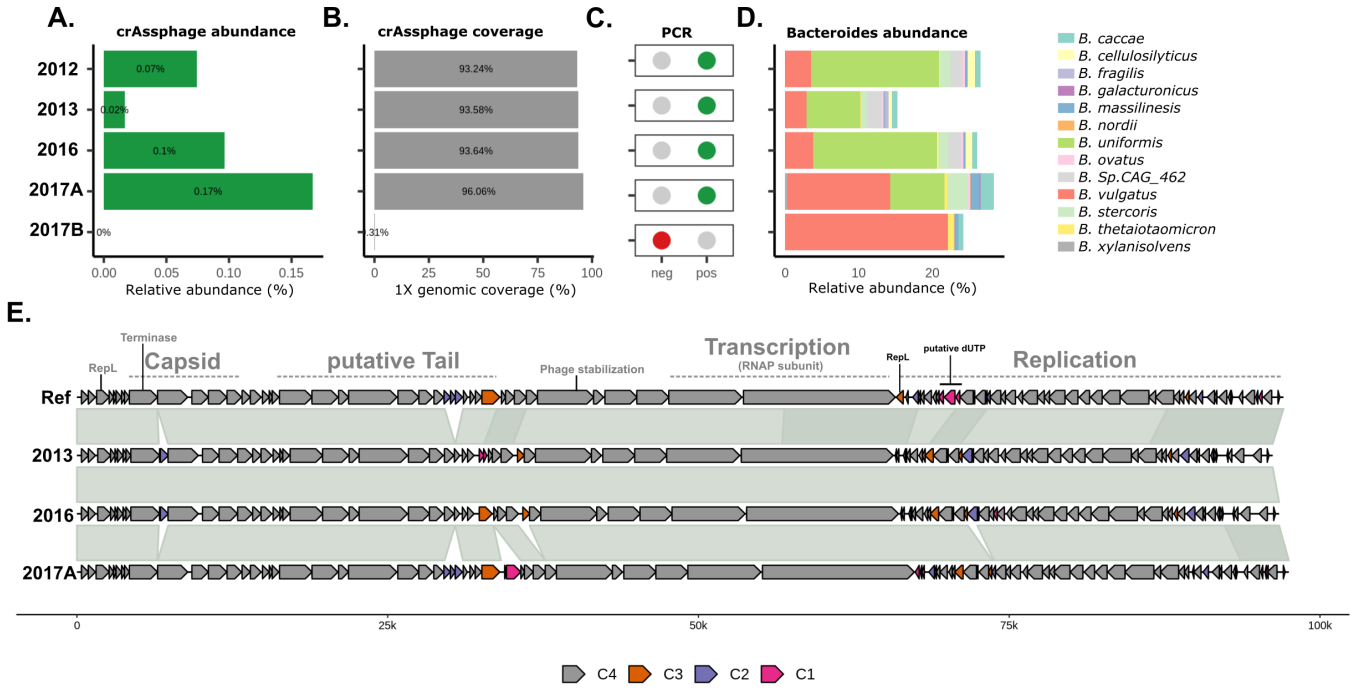

FIGURE 4.1: crAssphage dynamics in metagenomic samples from donor B. **A.** Relative abundance of crAssphage in longitudinal samples as a percentage of total reads. **B.** Percentage of reference crAssphage covered in each sample. **C.** PCR amplicon designed to detect crAssphage. **D.** Relative abundance of Bacteroides, potential crAssphage host, in each sample. **E.** Alignment and genome annotation of assembled crAssphage from metagenomic samples. The conserved genes that are common across these 4 genomes are shown in grey. Genes that are common among 3, 2, and 1 genome unique to samples are shown in color.

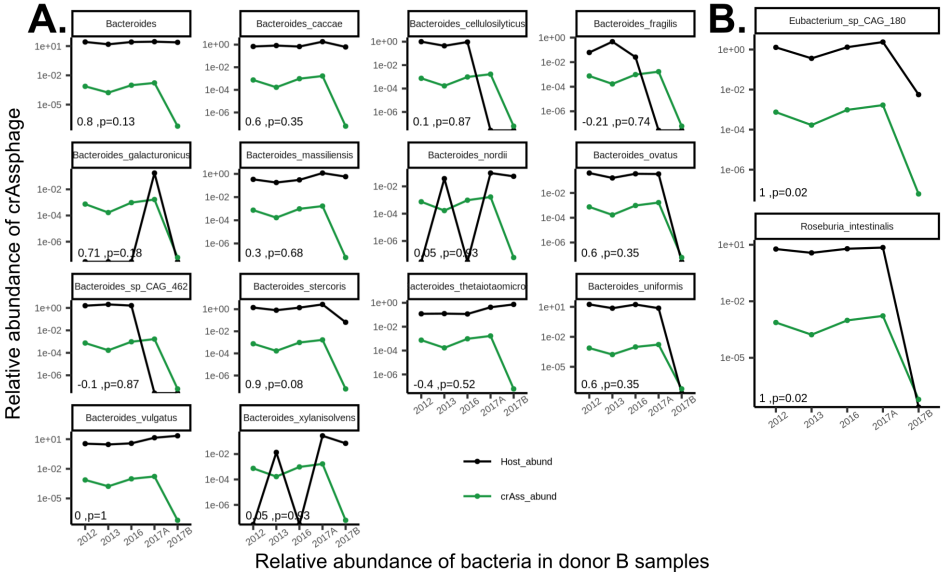

FIGURE 4.2: crAssphage−bacterial in donor. **A.** Relative abundance of crAssphage (x-axis) compared to relative abundance of Bacteroides genus (top left grid) and all the identified Bacteroides species. Green and black lines shows the relative abundace of crAssphage and bacteria in each sample respectively. **B.** *Eubaterium sp.* and *Roseburia intestinalis* that were significantly associated with crAssphage in donor B samples.

4.3.4 crAssphage contains homogeneous population but a variable strain in donor B

The metagenomic data represents not a single isolate but the population of the crAssphage present at that time. To investigate the population diversity at each time point, metagenomic reads from the longitudinal data were mapped to the *de novo* assembled crAssphage from the same sample. I observed relatively homogeneous crAssphage populations in each sample (Fig. 4.3). crAssphage in the 2016 and 2017A samples have shown the most and fewest observed SNPs, respectively. Here I define SNPs as positions with sequence variants at the population level that differ from the consensus (with a minimum threshold of 5% of the total reads at that position). As shown in Figure 4.3, the 2016 and 2017A samples contained 22 and 11 non-synonymous mutations, respectively, compared to the *de novo* assembly (consensus nucleotide) from

the same sample. Interestingly, the genome coverage was increased from 50X in 2016 to 250X in 2017A, but the number of detected SNPs reduced, showing that this crAssphage has a more homogeneous population.

To investigate how the crAssphage populations may change over time, the metagenomic reads from each time point were mapped to a single reference genome (the 2013 assembly).I found that donor B's crAssphage was stable from 2012 to 2016; however, the 2017A sample had a high number of sequence variants relative to the 2013 reference genome (2% of positions in the genome were different). This is consistent with displacement of the previous resident strain being replace by a different crAssphage strain (within subfamily alpha, cluster 1) in 2017A. This new phage completely disappeared within 7 months in the 2017B sample (Fig. 4.4).

4.3.5 crAssphage variability between individuals

To further examine population and strain variability between and within individuals, I used a cross-sectional metagenomic dataset from seven healthy donors. First, I mapped raw metagenomic reads from all seven samples (SHCM1-6 and SHCM15) to the *de novo* assembled crAssphage from donor B 2013. Only 2 out of 7 SHCM donors were positive, and the reference reference crAssphage was completely absent in the rest of these individuals (Fig. 4.5A) except SHCM1 that showed fragmented genome coverage, suggesting that the abundance of crAssphage for this donor was below the detection level or that it has a very different strain (Fig.4.5A). The aligned crAssphage reads from SHCM2 and SHCM4 had 926 and 873 non-synonymous mutations, respectively, versus the donor B 2013 assembly, suggesting these individuals carry a different strain. I also mapped shotgun reads from SHCM2 and SHCM4 to their own respective *de novo* assembled crAssphage genomes to investigate population variability in each sample. I found that the crAssphage populations in SHCM2 and SHCM4 were more heterogeneous than the donor B samples with 186 and 80 non-synonymous mutations (Fig. 4.5B).

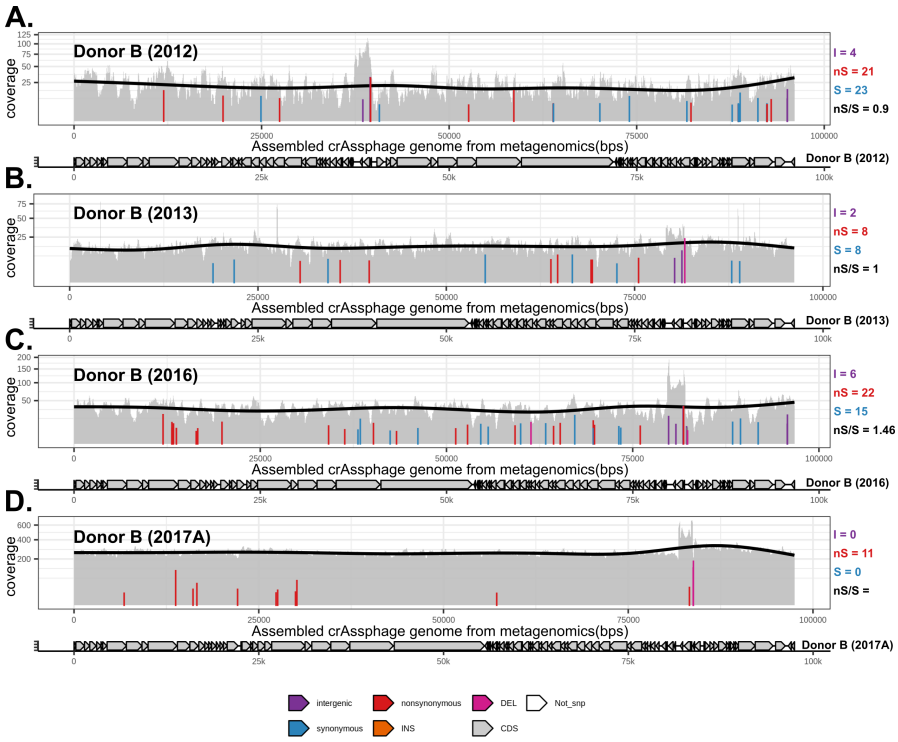

FIGURE 4.3: crAssphage populations in donor B samples. Comparison of detected SNPs compared to the consensus assembly of each sample as a measurement of crAssphage populations in four longitudinal samples from donor B. The x-axis shows the assembled crAssphage genome from metagenomics, and the y-axis shows crAssphage coverage. The SNP types are coloured as intergenic, synonymous, nonsynonymous. The length of each bar shows the frequency of that SNP.

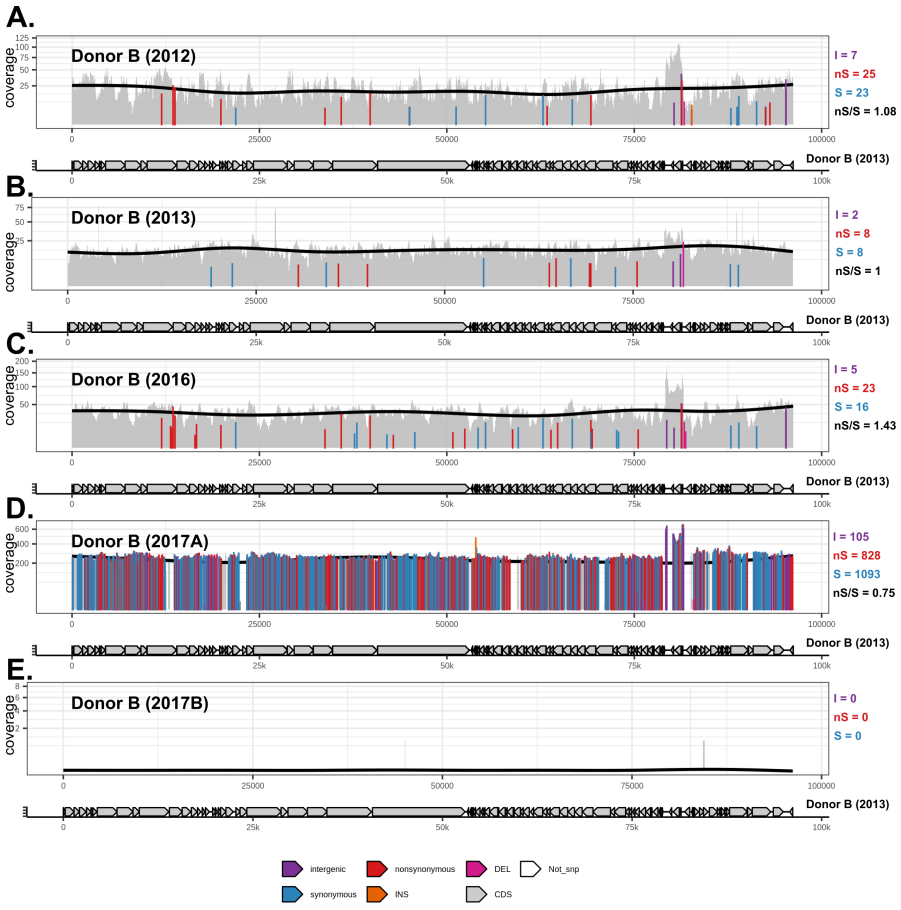

FIGURE 4.4: crAssphage variability in donor B samples. Comparison of detected SNPs in each sample compared to consensus assembly in 2013 sample. The x-axis shows the assembled crAssphage genome from metagenomics, and the y-axis shows crAssphage coverage. The SNP types are coloured as intergenic, synonymous, nonsynonymous. The length of each bar shows the frequency of that SNP.

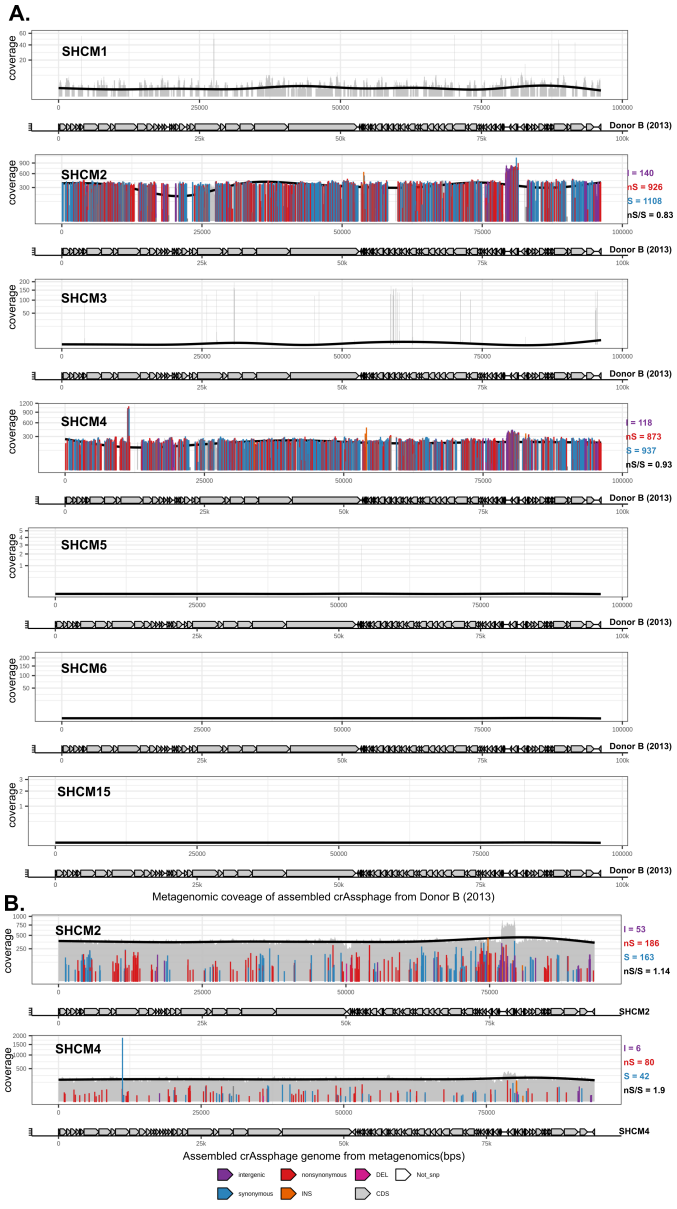

FIGURE 4.5: crAssphage variability in other healthy donors. Comparison of detected SNPs in each sample compared to consensus assembly in **A.** donor B 2013 and **B.** SHCM2 and SHCM4 samples. The x-axis shows the assembled crAssphage genome from metagenomics, and the y-axis shows crAssphage coverage. The SNP types are coloured as intergenic, synonymous, nonsynonymous. The length of each bar shows the frequency of that SNP.

4.3.6 Detection of crAssphage engraftment requires population information

I have investigated paired metagenomic assemblies (pre-and post-FMT) from 11 UC patients (22 samples) who received FMT from donor B. crAssphage was only present in 3 out of 22 samples. I was able to identify crAssphage in pt80 only post-FMT, and pt84 was crAssphage positive in both pre-and post-FMT samples. In order to track population variabilities, I mapped metagenomic reads from these three samples to the *de novo* assembly from the same sample and compared them to the donor B 2013 assembly. The post-FMT sample from pt80 showed homogeneous crAssphage population with a modest increase in variability compared to donor B (Fig. 4.6A,B). However, pt-84's post-FMT variability was almost identical to the sample pre-FMT with 26 out 32 variable site SNPs shared(Fig. 4.6C,D).

To test whether the crAssphage detected post-FMT was engrafted from donor B, I mapped metagenomic reads from these two patients to the *de novo* assembled crAssphage in donor B (2013). I found that the crAssphage strain post-FMT in pt80 was different from the donor B with 849 nonsynonymous and 977 synonymous mutations, suggesting that the detected crAssphage was not engrafted from donor B (Fig. 4.7A,B). The crAssphage strain post-FMT in pt84 was identical to the one pre-FMT based on the genomic gaps and number of observed mutations between these two samples (Fig. 4.7C,D).

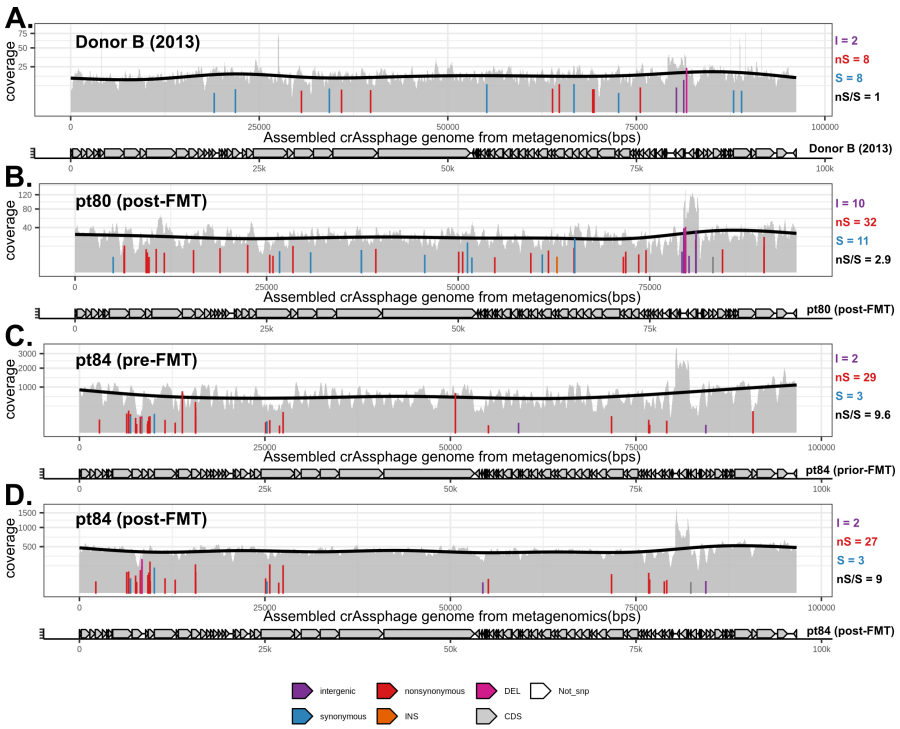

FIGURE 4.6: crAssphage population in UC patients pre-and post-FMT treatment. Comparison of detected SNPs as a measurement of crAssphage populations in pt80 that was crAssphage positive only post-FMT and pt84 that contains crAssphage both pre-and post-FMT. For each sample, the x-axis shows the assembled crAssphage genome from metagenomics, and the y-axis shows crAssphage coverage. The SNP types are coloured as intergenic, synonymous, nonsynonymous. The length of each bar shows the frequency of that SNP.

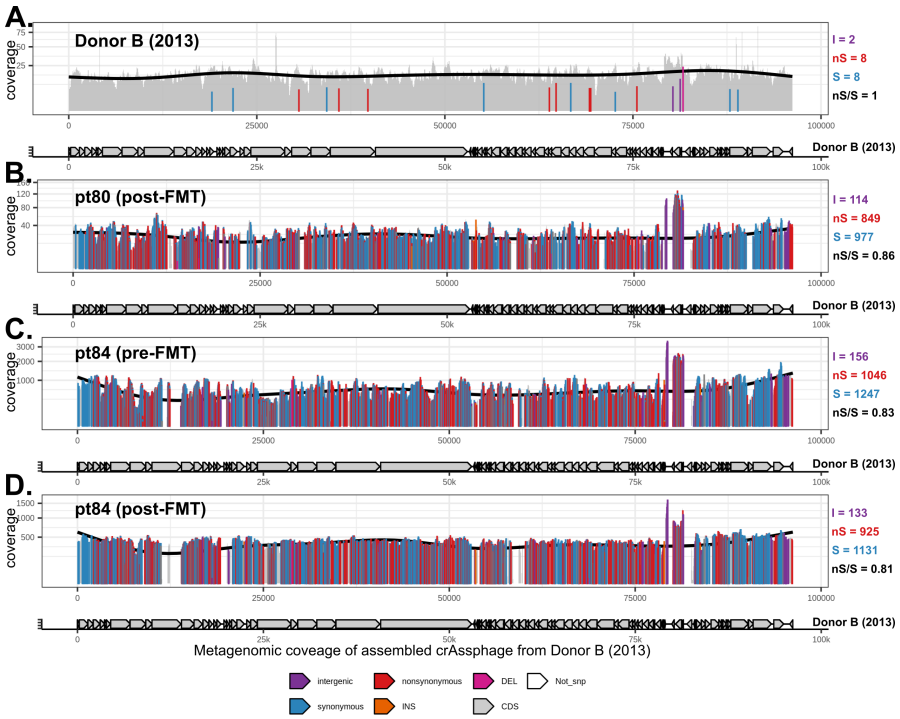

FIGURE 4.7: crAssphage variability in UC patients pre-and post-FMT treatment. Number of detected SNPs in pt80 and pt84 compared to consensus assembly in donor B 2013 sample. The x-axis shows the assembled crAssphage genome from metagenomics, and the y-axis shows crAssphage coverage. The SNP types are coloured as intergenic, synonymous, nonsynonymous. The length of each bar shows the frequency of that SNP.

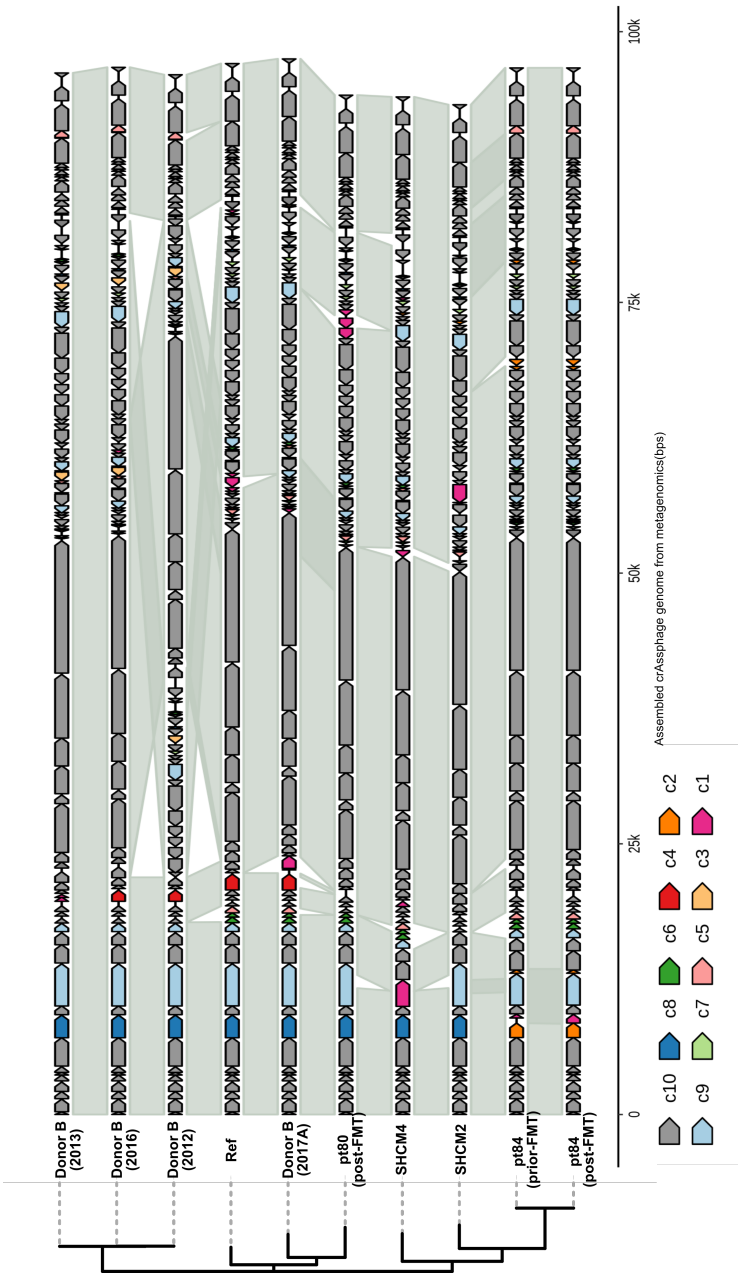

FIGURE 4.8: Circular whole-genome alignment of crAssphage assembled from metagenomic samples.

4.3.7 crAsSNPer: a method to detect crAssphage engraftment using metagenomics

I showed that accurate detection of crAssphage engraftment requires SNP information. However, the number of reported SNPs depends on the sequencing depth and crAssphage abundance (copy number) in a sample. To address these challenges, I developed a method (crAsSNPer) for accurate detection of crAssphage engraftment in samples with variable depths of sequencing and crAssphage abundance. crAsSNPer conducts a bootstrap by randomly sampling crAssphage reads from a metagenomic sample using the lowest mean depth of crAssphage in a sample from the same dataset and re-calculates the total number of SNPs. The user can change the number of iterations for sampling with replacement, but the default crAsSNPer subsamples the data twenty times. crAsSNPer uses a linear model to find the upper and lower confidence intervals for the expected crAssphage SNP frequencies for an individual (multiple samples) or sample (see Methods).

To test the performance of crAsSNPer, I used a viral metagenomic dataset (PRJNA446038) from a previously published FMT study for patients with rCDI (Draper et al. 2018). I assessed longitudinal samples from a donor (D3, 16 samples) who was crAssphage positive, as well as samples from patients who received FMT from donor D3. For each patient, one sample was collected before FMT, and seven longitudinal samples were collected post-FMT up to one year (54 samples in total) (Draper et al. 2018). I also investigated the data from a single FMT recipient (P7) who received FMT from donor D1 (crAssphage negative) but became crAssphage positive two months after FMT (4 samples). The sample from P7 shows expected SNP frequencies from a different crAssphage strain and can be used as a negative control for accurate detection of crAssphage engraftment in patients who received FMT from donor D3. A *de novo* reference genome with a total length of 92kb was assembled by co-assembly of the F0 and F1 samples from donor D3 and all the samples from the donor and patients were compared to this reference genome.

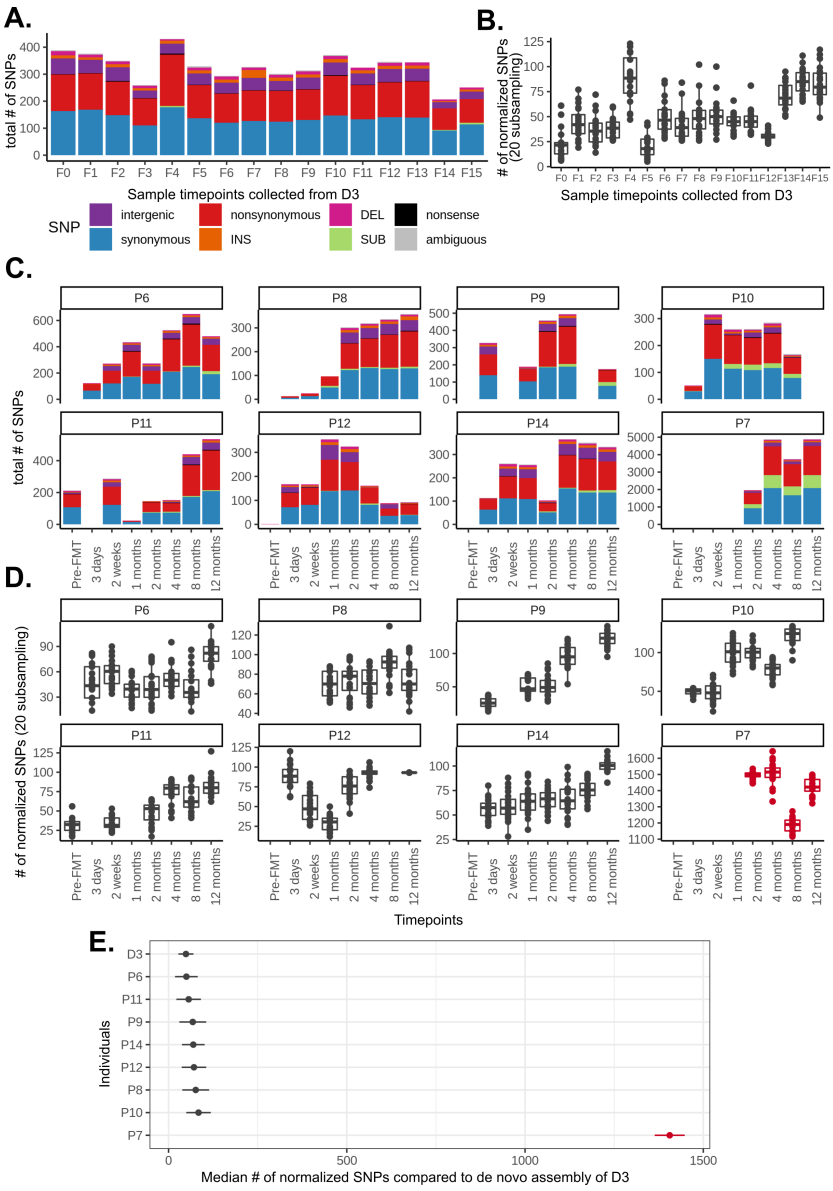

FIGURE 4.9: (Caption next page.)

FIGURE 4.9: Evaluation of the crAsSNPer using a publicly available viral metagenomic dataset. **A.** crAssphage variation in longitudinal samples collected from donor D3 (F0−F15) compared to the *de novo* assembled crAssphage from D3. The y-axis shows the total number of SNPs identified in each sample. The colours in each bar shows the SNP types. **B.** crAssphage variation in donor D3 using the crAsSNPer pipeline. The y-axis shows a normalized number of SNPs (20 subsampling) for each sample on the x-axis. **C.** crAssphage variation in patients who received FMT from donor D3. The y-axis shows the total number of SNPs identified in each sample of patients. Longitudinal samples collected before and post-FMT are shown on the x-axis. Patients P6, P8, P9, P10, P11, P12, and P14 received FMT from donor D3, and patient P7 received FMT from a different donor who was crAssphage negative. **D.** crAssphage variation in patients who received FMT from donor D3 (black) and patient P7 (red) using the crAsSNPer pipeline. The y-axis shows a normalized number of SNPs (20 subsampling) for each sample on the x-axis. **E.** Comparison of the identified crAssphage in patients who received FMT from donor D3 (black) versus patient 7. Each dot shows a median number of normalized SNPs on x-axis for all the samples for each individual with 20 subsampling. The upper and lower confidence intervals are shown around each dot.

Our result showed that donor D3's crAssphage is moderately stable over time, with F4, F14, and F15 samples showing slightly increased SNP frequencies compared to the reference genome (Fig. 4.9A,B). The samples collected from D3's patients post-FMT (P6, P8, P9, P10, P11, P12, and P14) showed variable SNP frequencies over time based on the total number of reported SNPs (Fig. 4.9B), and normalized number of SNPs with 20X subsampling (Fig. 4.9D). Interestingly, the SNP frequencies increased post-FMT over time compared to donor D3, particularly in patients P9, P10, P11, and P14 (Fig. 4.9D). However, these variations were close to the expected boundaries of donor D3 and significantly different (est=1357, $p < 0.0001$) compared to the median number of SNPs in patient P7 (Fig. 4.9E). These results confirmed Draper et al. 2018 findings suggesting the engraftment of donor D3's crAssphage in patients post-FMT.

Next, I applied crAsSNPer to the data from donor B and their FMT recipients. I used donor B's *de novo* assembled crAssphage from 2013 as a reference. Consistent with our previous results (section 4.3.4), crAsSNPer also showed that donor B's crAssphage

was stable from 2012 up to 2016 but contained a significantly different (est=1957.40, p< 0.0001) median number of SNPs in the sample collected from 2017A, suggesting strain replacement (Fig.4.10A, B). I also showed that pt84, who was only positive for crAssphage post-FMT, contains significantly different (est=1785.65, p< 0.0001) crAssphage strain based on the median number of SNPs (Fig.4.10B). Similarly, patient 84's crAssphage from pre-and post-FMT showed a significantly different (pre-FMT; est=1382, p< 0.0001, post-FMT; est=1958.35, p< 0.0001) median number of SNPs at the level expected in other individuals (SHCM2, SHCM4).

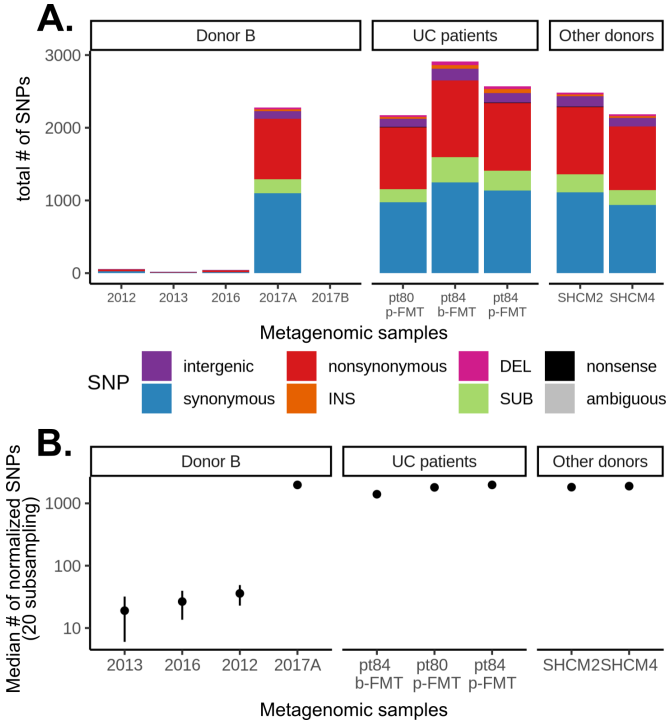

FIGURE 4.10: Accurate detection of donor B's crAssphage post-FMT using the crAsSNPer. **A.** crAssphage variation in samples collected from donor B, UC patients, and two other healthy individuals compared to the *de novo* assembled crAssphage from donor B (2013 sample). The y-axis shows total number of SNPs identified in each sample. The colours in each bar shows the SNP types. **B.** Comparison of the identified crAssphage in donor B, patients who received FMT from donor donor B (pt80 and pt84), and two healthy donors (SHCM2, SHCM4). Each dot shows a median number of normalized SNPs on y-axis for each sample (n=20 subsampling). The upper and lower confidence intervals are shown around each dot.

4.4 Discussion

crAssphage is a highly abundant bacteriophage (Dutilh et al. 2014) that has co-evolved with humans for millions of years (Edwards et al. 2019). This phage belongs to a family of crAss-like phages consisting of at least four subfamilies (Guerin et al. 2018). crAssphage is globally distributed and locally clustered within individuals (Edwards et al. 2019). It has been shown that crAssphage is stable and can transfer between individuals within a household (Siranosian et al. 2020) or via FMT (Draper et al. 2018). However, the temporal stability of this bacteriophage within and between individuals is not well understood. I asked whether individuals can still show temporal crAssphage changes and whether these changes are related to the transferability of crAssphage between individuals. Here I have reported a case of variable crAssphage in a single healthy individual. I have argued that given these temporal strain replacements, accurate detection of crAssphage transfer requires high-resolution SNP information, as these changes are not captured in PCR data alone.

A stable crAssphage strain in donor B was replaced by a closely related strain with an increased relative abundance, and subsequently disappeared in this individual within less than five months. The replacement strain in donor B is more similar to the NCBI crAssphage reference, and the difference is in the presence/absence of genes as well as SNP frequencies. The variability in the relative abundance of crAssphage in donor B potentially reflects changes in their bacterial host. It has been shown that *Bacteroides sp.* are potential crAssphage hosts, and using enrichment-based techniques, the host for one member of the crAss family, crAss001, was confirmed to be *Bacteroides intestinalis*(Shkoporov et al. 2018). Although changes in the relative abundance of *B. ovatus* and *B. uniformis* were similar to crAssphage, I have not observed any significant correlation between *Bacteroides sp.* and crAssphage in donor B. The increased relative abundance of *B. vulgatus* in the 2017B sample may suggest that a single none

crAssphage *Bacteroides sp.* host took over the microbial community and out-competed the crAssphage host.

Previously, it was shown that crAssphage could transfer from healthy individuals to patients with rCDI and stay stable for up to one-year (Draper et al. 2018; Siranosian et al. 2020). Because *C. difficile* is over-represented in the intestinal microbiota in patients with rCDI, crAssphage hosts are potentially diminished in these patients. I argue that in a more complex microbiome in which a single bacterium does not dominate the community (UC patients), a pre-existing crAssphage prior to FMT would compete with the one present in the donor. crAssphage polymerase chain reaction (PCR) data from UC patients pre-and post-FMT suggested successful transfer of this phage post-FMT. However, using high-resolution SNP analysis from metagenomic data, I showed that a completely different strain was, in fact, present in patient 84 post-FMT. It is also possible that strains at very low levels will not be detected in donor samples but expand in the recipients, which can be a caveat for any engraftment detection. These results suggest that accurate detection of crAssphage engraftment post-FMT requires whole genome population information.

crAsSNPer can detect crAssphage transfer despite variation in sequencing depth and crAssphage copy number in the metagenomic samples. Using this pipeline, I have confirmed Draper et al. 2018 findings suggesting the engraftment of donor D3's crAssphage in rCDI patients post-FMT. crAsSNPer was able to show crAssphage SNP variation within and between individuals. Most importantly, using this pipeline, I showed that donor D3's crAssphage is substantially different from that in patient P7, who was crAssphage positive but did not receive FMT from donor D3.

Chapter 5

Efficacy of antimicrobials versus placebo in addition to FMT in patients with ulcerative colitis

5.1 Introduction

Ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD) that is characterized by colonic mucosa inflammation. The etiology of UC is unknown but it is suspected to be an immune response to altered intestinal microbiota in predisposed individuals. Fecal microbiota transplantation (FMT) — transfer of stool content from healthy, screened individual to patients — is a proposed treatment for UC. FMT is an existing therapy for patients with recurrent- *Clostridioides difficile* infection (rCDI), but its efficacy against UC remains an open question. Previous randomized controlled trials (RCTs) have shown that FMT can alter colonic microbiota by microbial engraftment — the colonization of donor's microbiota in patients post-FMT — and that these changes are associated with remission in a subset of UC patients (Chapter 3).

The microbiology of IBD is complex, as the active disease will alter the microbiome. Identifying which features of the changing microbiota are cause or consequence of inflammation has been challenging to resolve. It is not yet clear if specific pathogens drive inflammation; however, a few studies have suggested that enteric pathogens are involved in disease complications (Petersen et al. 2009; Mirsepasi-Lauridsen et al. 2016; Axelrad et al. 2018). Enteric infection is frequently seen in UC patients, but little is known regarding the distribution and genomic variability of those pathogens (Axelrad et al. 2018). The efficacy of antibiotics in treating UC flare-ups suggests that eliminating or reducing some bacterial pathogens may result in disease improvement (Khan et al. 2011).

A combination of antimicrobial and FMT therapies can potentially enhance FMT outcome. It has been implied that pretreatment with antibiotics increased the efficacy of FMT (Ishikawa et al. 2017; Keshteli et al. 2017), but it is not clear whether that improvement is associated with microbial changes or engraftment post-FMT. In this study, we report the first RCT of antibiotics versus placebo in combination with FMT

for active UC patients. We investigated whether these therapies are associated with microbiome changes post-FMT.

5.2 Methods

5.2.1 Study design

A randomized placebo-controlled trial was conducted at McMaster University to evaluate whether adding an antimicrobial cocktail prior to FMT increases the remission rate in patients with ulcerative colitis. The recruited patients received two antibacterial agents (metronidazole 500mg, doxycycline 100mg) and an antifungal (terbinafine 250mg), or placebo once daily for two weeks prior to FMT. Within 1-3 days post completion of their antimicrobial/placebo course, patients received FMT enemas twice weekly for eight weeks.

5.2.2 Study population, clinical outcome, and sample collection

Seventy-five patients were assessed for trial eligibility. Active UC patients — Mayo score > 3 and endoscopic Mayo score > 0 — who were ≥ 18 years were included in the trial. Exclusion criteria were defined as severe UC requiring hospitalization, *Clostridium difficile* infection, severe comorbid medical illness, antibiotic therapy in the last 30 days, increase in medical therapy for UC in the last 12 weeks, and any condition that the treatment may pose a health risk. The trial's primary outcome was defined as a Mayo score < 3 with an endoscopic Mayo score $= 0$ at the end of the trial (week 9). Fecal samples were collected from each patient at baseline, after two weeks of antimicrobial treatment, and post-FMT at week 9. A sample was taken from every batch of FMT slurry from each donor.

5.2.3 Genomic DNA extraction and 16S rRNA amplicon sequencing

Genomic DNA was extracted using the MagMAX Express 96-Deep Well Magnetic Particle Processor from Applied Biosystems with the Multi-Sample kit (Life Technologies # 4413022) with the addition of a bead beating step as described in Chapter 2. Purified DNA was used to amplify the v34 region of the 16S rRNA gene by PCR. 50 ng of DNA was used as template with 1U of Taq, 1x buffer, 1.5 mM MgCl2, 0.4 mg/mL BSA, 0.2 mM dNTPs, and 5 pmoles each of 341F (CCTACGGGNGGCWGCAG) and 806R (GGACTACNVGGGTWTCTAAT) with Illumina adapters and barcodes, as described in Bartram et al. 2011. The reaction was carried out at 94C for 5 minutes, 5 cycles of 94C for 30 seconds, 47C for 30 seconds and 72C for 40 seconds, followed by 25 cycles of 94C for 30 seconds, 50C for 30 seconds and 72C for 40 seconds, with a final extension of 72C for 10 minutes. Resulting PCR products were visualized on a 1.5% agarose gel. Positive amplicons were normalized using the SequalPrep normalization kit (ThermoFisher#A1051001) and sequenced on the Illumina MiSeq platform at the McMaster Genomics Facility.

5.2.4 16S rRNA gene amplicon sequencing processing pipeline

Reads were processed using DADA2 (Callahan et al. 2016). First, Cutadapt (Martin 2011) was used to filter and trim adapter sequences and PCR primers from the raw reads with a minimum quality score of 30 and a minimum read length of 100bp. Sequence variants were then resolved from the trimmed raw reads using DADA2. DNA sequence reads were filtered and trimmed based on the quality of the reads for each Illumina run separately, error rates were learned and sequence variants were determined by DADA2. Sequence variant tables were merged to combine all information from separate Illumina runs. Bimeras were removed and taxonomy was assigned using the SILVA database version 1.3.8 Quast et al. 2012.

The ASV table was rarefied to the lowest read count to measure microbial diversity

in each sample and the Shannon values were estimated based on the rarefied ASV table using the phyloseq (McMurdie and Holmes 2013) package. A custom function was written in R 4.2.0 to parse result tables and visualize the sample's diversity using tidyverse 1.3.1. In order to test the diversity difference across group variables, the desired samples were selected for each variable, and a linear mixed-effect model was fitted with sampling timepoint as the fixed effect and patient ID as random effect using lme4 and lmerTest (Bates et al. 2014) packages in R 4.2.0.

To visualize sample distances (beta-diversity), two different distance metrics were used; 1) Bray–Curtis based on the relative abundance of ASVs 2) Aitchison distance based on centered log-ratio (CLR) transformed of ASV counts. For Bray-Curtis distance, the ASV table was transformed to relative abundance, and the distances among samples were visualized with a Principal Component Analysis (PCoA) using a custom function incorporating phyloseq (McMurdie and Holmes 2013) in R 4.2.0. For Aitchinson distance, the ASV counts were transformed to the centered log-ratio (CLR) using microbiome v.1.12.0 and visualized via Principal Component Analysis (PCA). A permanova test on Bray–Curtis and Aitchison distances to measure microbial shift between time points using a custom function incorporating the ape and vegan packages (Oksanen et al. 2013) in R 4.2.0.

ANCOMBC (Lin and Peddada 2020) was used to identify the differentially abundant ASVs between time points for each group. ASVs with adjusted p-values < 0.05 were visualized using a custom function in R 4.2.0. To compare engrafted ASVs among donors, donor-specific ASVs — the ASVs that were present in at least one sample from one donor — were selected. Their relative abundance was compared between baseline and post-FMT samples for each patient. ASVs with a relative abundance of 0 at baseline and $\geq 0.1\%$ post-FMT were defined as engrafted. We then compared the number of engrafted ASVs across an increasing number of patients who received donor B vs. donor

M1 FMTs. Since donors B and M1 were used for different numbers of patients, we randomly sampled 15 patients from each donor (with 100 re-sampling) and compared the number of engrafted ASVs across an increasing number of patients to estimate confidence intervals. Two custom functions were written in R 4.2.0 to identify engrafted ASVs and perform the permutations as described above. All the code used above is available at https://github.com/SShekarriz/UCFMT2

5.3 Results

A previous RCT by our team demonstrated that FMT improved UC over placebo and induced endoscopic remission in 24% of subjects (Moayyedi et al. 2015). A second RCT has been completed to determine in pretreatment of subjects with antibiotic would improve efficacy of FMT. The treatment group received antibiotics (metronidazole 500mg, doxycycline 100mg, terbinafine 250mg) for two weeks followed by FMT therapy (twice weekly for eight weeks). 63 out of 75 UC patients screened were eligible for the trial (see Methods). 31 and 32 were randomly assigned to receive antibiotic therapy and placebo intervention, respectively. Fecal samples were collected at baseline, after antibiotic therapy, and during the last week of FMT (Fig. 5.1). 4 patients (three from the antibiotic group and one on placebo intervention) withdrew before the completion of the trial. However, their baseline and post-antibiotic sampes were included in this analysis.

17 of the 28 (61%) patients who received antimicrobial pretreatment showed partial improvement — defined as $< 33\%$ reduction in partial Mayo clinic score — versus 20 out 31 (65%) who received placebo pretreatment. Endoscopy post-FMT was not completed for three patients who were randomized to antibiotic therapy and four patients on placebo intervention due to hospital closure during the Covid19 pandemic. 7 of the 25 (28%) patients with full week nine endoscopy from the antimicrobial group went into clinical remission — defined as a Mayo score < 3 with an endoscopic Mayo score of 0 at the end of the trial — compared to 9 of the 27 (33%) of those on placebo pretreatment. Three samples from patients who received antibiotic therapy (one baseline and two post-FMT samples) and five samples from the placebo group (two baseline, one post-antibiotic, and two post-FMT) were not available for sequencing (Fig. 5.1).

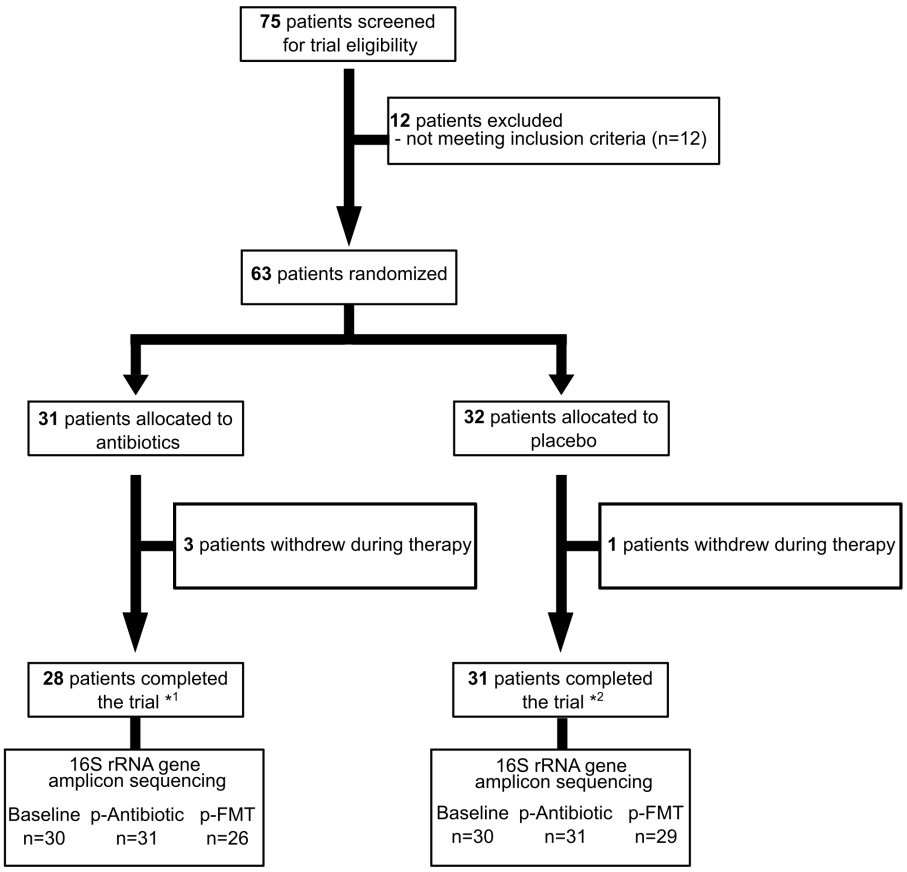

FIGURE 5.1: Flow chart of enrolled patients and fecal samples collected for 16S rRNA gene amplicon sequencing. The mucosal healing at the end of the study (week 9) was not assessed for 3 patients randomized to antibiotic therapy (*1) and 4 patients on placebo intervention (*2) due to hospital closure during the Covid19 pandemic. Fecal samples collected at Baseline, post-antibiotic (p-Antibiotic), and post-FMT (p-FMT) were used for 16S rRNA gene amplicon sequencing.

5.3.1 Pretreatment with antimicrobials alters the microbiome but does not induce a greater charge by FMT.

To test whether antibiotic therapy before FMT would increase the diversity of microbiota post-FMT, we have identified amplicon sequence variant (ASV) in all the fecal samples collected at baseline, post-antibiotic, and post-FMT using 16S rRNA gene amplicon sequencing. We used the Shannon diversity index to measure the alpha diversity in each sample. As we expected, the mean difference between baseline and post-antibiotic samples was significantly bigger in patients who received antibiotics compared to placebo (LMM, est=0.8, p=0.0001; Fig. 5.2A). The mean difference between post-antibiotic and post-FMT samples was significantly smaller in the antibiotic group compared to the placebo (LMM,est=-0.6,p=0.002; Fig. 5.2A). However, the mean difference between baseline and post-FMT was not significantly different (LMM, est:0.14, p=0.4; Fig. 5.2A) between these two groups.

In order to ask whether pretreatment with antimicrobial therapy induces microbiota community-wide shift post-FMT, Bray-Curtis and Aitchison distances were calculated and compared pairwise samples within each patient. We observed that the mean distance between baseline and post-antibiotic was significantly greater in patients who received antibiotics compared to placebo (Bray-Curtis: Anova, antibiotic=0.71,placebo=0.51, se=0.04,p=0.0008; Fig. 5.2C and Aitchison: antibiotic=71, placebo=55,se=2.4,p=2.9e-05; Fig. 5.2D), suggesting a microbial community change after two weeks of antibiotic treatments. However, the mean distance between baseline and post-FMT was not significantly different in antibiotic group compared to placebo intervention (Bray-Curtis: Anova, antibiotic=0.67,placebo=0.68,se=0.03,p=0.8; Fig. 5.2C and Aitchison: antibiotic=80, placebo=76, se=2.4,p=0.2; Fig. 5.2D).

Next, we asked whether there were any ASVs that differentially abundant in antibiotic group compared to placebo across different time-points. ANCOMBC (Lin and

Peddada 2020) was used with interactions effects between sample's time and patients interventions to normalize and identify the different ASVs (p-adjust < 0.05, see Methods). Figure 5.3B compares these ASVs for patients who received placebo versus antimicrobial interventions prior to FMT. Most of these ASVs shared bacterial families across antibiotic and placebo groups. Most notably, the patients who received antibiotic showed increased abundance of Enterococcaceae and reduction of multiple ASV belonging to Peptostreptococcaceae in their post-antibiotic samples. The abundance of Enterococcaceae, Prevotellaceae, and Sutterellaceae ASVs reduced and Peptostreptococcaceae increased post-FMT in patients who received antibiotic treatment compared to placebo.

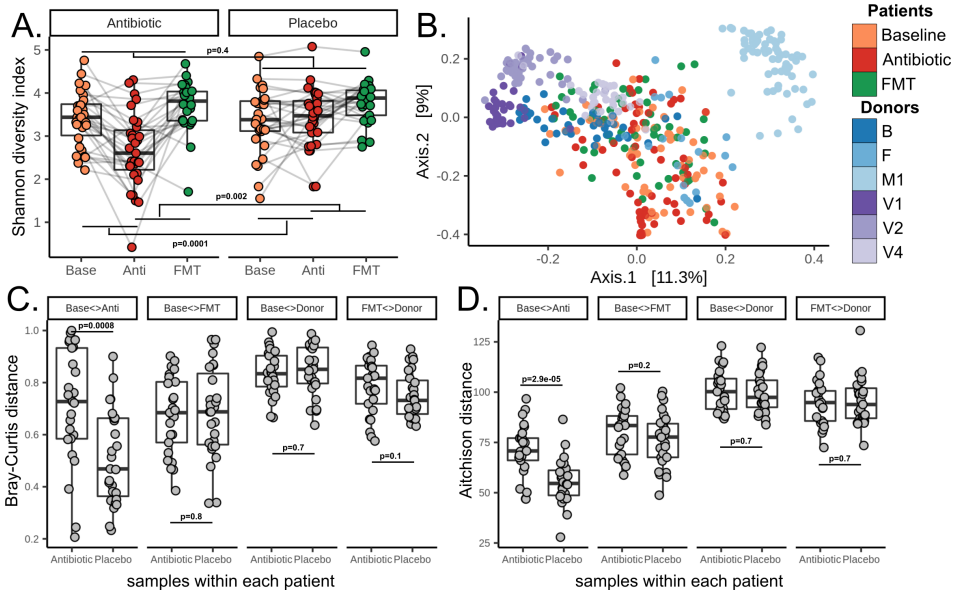

FIGURE 5.2: Comparison of the antibiotic versus placebo treatment prior to FMT therapy. **A.** The Shannon alpha diversity metric for samples collected from patients baseline (Base), post-antibiotc (Anti), and post-FMT (FMT). The left and right facets shows patients who received antibiotic and placebo treatments prior to FMT. **B.** PCoA of Bray-Curtis, beta diversity, distances between all samples. Pairwise Bray-Curtis (**C.**) and Aitchison (**D.**) distances between samples within patients who received antibiotic or placebo pretreatments. Distances are measured between baseline vs. post-antibiotic (Base<>Anti), baseline vs. post-FMT (Base<>FMT), baseline vs. donor (Base<>Donor), and post-FMT vs. donor samples. The last donor sample for each patient was used to calculate these distances.

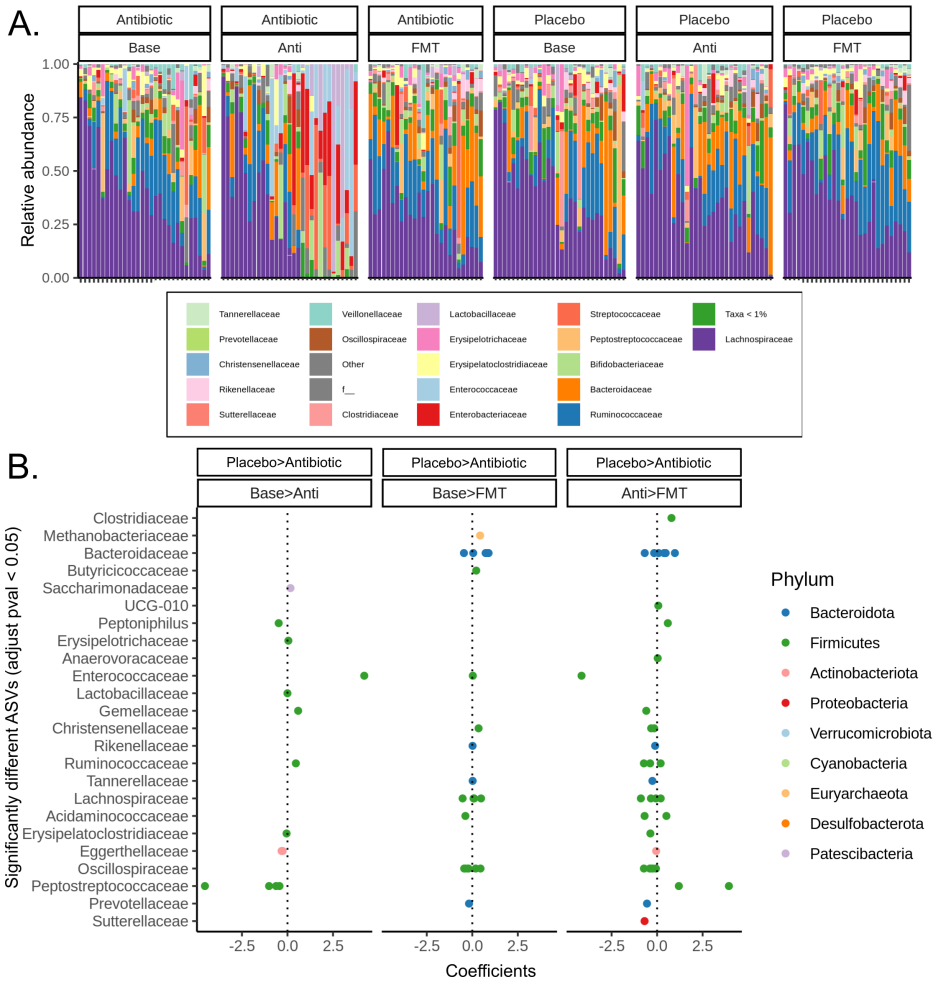

FIGURE 5.3: Taxonomic composition of fecal samples in antibiotic versus placebo treatment. **A.** Relative abundance of bacterial families in samples collected at baseline (Base), post-antibiotic (Anti), and post-FMT (FMT) in patients who either received antibiotic or placebo interventions. **B.** Significantly different ASVs between placebo and antibiotic (Placebo>Antibiotic) interventions compared across baseline versus post-antibiotic (Base>Anti), baseline versus post-FMT (Base>FMT), and post-antibiotic versus post-FMT (Anti>FMT) samples. x-axis shows a natural log of coefficients.

5.3.2 Microbial shift is not specific to patients who responded to treatments

In order to explore whether the clinical outcome was associated with microbial change after antimicrobial pretreatment, microbial composition of baseline and post-FMT samples was compared in patients who showed clinical remission (responders; n=16, 27%) versus those who did not respond to the treatment (non-responders; n=36,61%) at the end of the trial. 7 (12%) patients who completed the trial but not the endoscopy at the end of the trial were excluded from our analysis.

Shannon index based on identified ASVs was used to measure alpha diversity in each sample. The mean difference between baseline and post-FMT samples was not significantly different in patients who responded to therapy compared to non-responders (LMM, est=0.32, p=0.15; Fig. 5.4A). Bray-Curtis and Aitchison distances were used to test whether remission resulted in microbial community shift. No significant difference between responder and non-responder patients in inducing more microbial community shift was observed using pairwise samples within each patient (Anova, Bray-Curtis: NoRes=0.7,Res=0.75, se=0.5,p=0.7 Fig. 5.4C and Aitchison: NoRes=84,Res=80, se=5,p=0.3 Fig. 5.4D), suggesting that these change are not detectable at the community-level.

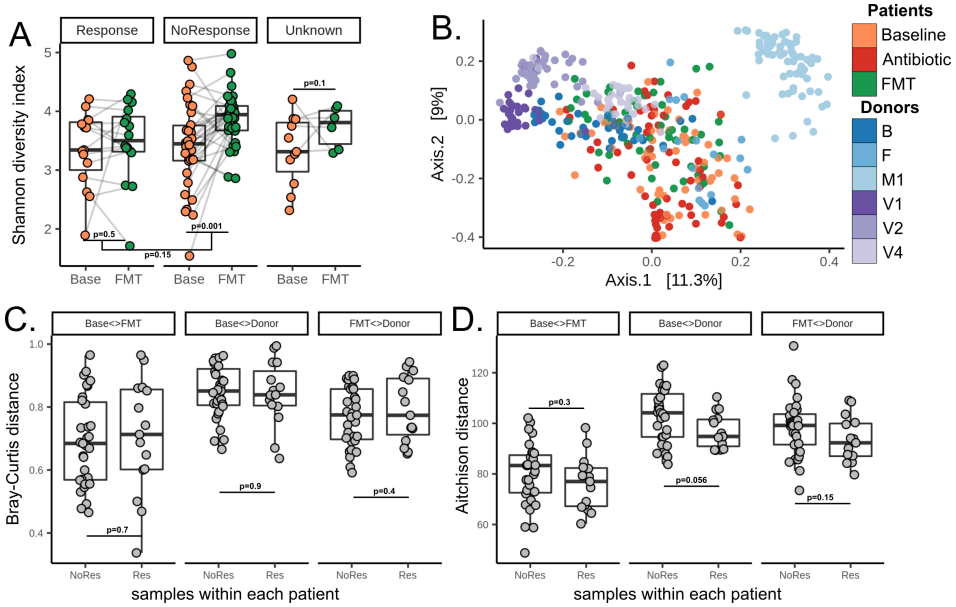

FIGURE 5.4: Microbial change post-FMT is not associated with clinical outcome. **A.** The Shannon index for samples collected at baseline (Base) and post-FMT (FMT) in patients who went to remission (Response), did not respond to treatments (NoResponse), and those who completed the trail but not the final endoscopy (Unknown). **B.** PCoA of Bray-Curtis distances between all samples. Pairwise Bray-Curtis (**C.**) and Aitchison (**D.**) distances between samples within patients who were non-responder (NoRes) and responder (Res) at the end of the trial. Distances are measured between baseline and post-FMT (Base<>FMT), baseline and donor (Base<>Donor), and post-FMT and donor (FMT<>Donor) samples. The last donor sample for each patient was used to calculate these distances.

5.3.3 Donor affects microbial change post-FMT.

To examine whether the donor can affect the microbial shift post-FMT, donors B and M1, who donated fecal slurries to 19 and 24 patients, respectively, were compared. The Shannon index was used to calculate alpha diversity for paired samples collected at three time-points. We observed that the mean difference between baseline and post-FMT samples was significantly smaller in patients received FMT from donor M1 compared to donor B (LMM,est=-0.5,p=0.01). However, patients who received donor B FMT were significantly less diverse than donor samples at baseline (Anova, Base=3.3, B=3.9,se=0.11,p=1.9e-05; Fig. 5.5A) and became similar to donor B's post-FMT (Anova, FMT=3.9, B=3.9,se=0.05,p=0.9; Fig. 5.5A). However, the mean Shannon index for donor M1's patients was not different from the donor samples, neither at baseline nor post-FMT (Anova, Base=3.4,M1=3.5,se=0.04,p=0.2; FMT=3.6, se=0.04, p=0.3; Fig. 5.5A). More interestingly, less variability in Shannon values post-FMT compared to baseline or post-antibiotic samples was observed (Fig. 5.5A).

Bray-Curtis distance was used to test whether donor B or M1 can induce microbial community shift post-FMT (Fig. 5.5C-F). Our results showed that the microbial community in patients who received FMT from donor B was changed post-FMT (R2=3.2%,p=0.01; Fig. 5.5D), while those who received FMT slurries from donor M1 (R2=1.4%,p=0.36; Fig. 5.5E) or V4 (R2=9%,p=0.04; Fig. 5.5F) did not show a significant difference post-FMT.

Comparing samples within each patient, there was no significant difference in community change post-FMT between donor B and M1 patients (Anova Bray-Curtis: B=0.7,M1=0.7, se=0.1,p=0.9; Fig. 5.5G and Aitchison: B=80,M1=80, se=0.1,p=0.4; Fig. 5.5H). The patients who received slurries from donor B were significantly more similar to the donor sample post-FMT (Anova Bray-Curtis: Base-B=0.87, FMT-B=0.78, se=0.02, p=0.01; Fig. 5.5G) but not significantly different in patients received FMT from

donor M1 (Anova Bray-Curtis: Base-B=0.8, FMT-B=0.75, se=0.02, p=0.09; Fig. 5.5G).

A group of ASVs were found to be differentially abundant (p-adjust < 0.05, see Methods) in patients who received FMT from donor B compared to donor M1 across different time-points. As shown in figure 5.6B, ASVs that were belonged to Prevotellaceae, Anaerovoracaceae, Christensenellaceae, Ruminococcaceae, Bacteroidaceae, Anaerococcus, Acidaminococcaceae, Oscillospiraceae, Lachnospiraceae, Rikenellaceae were increased in donor B patients post-FMT.

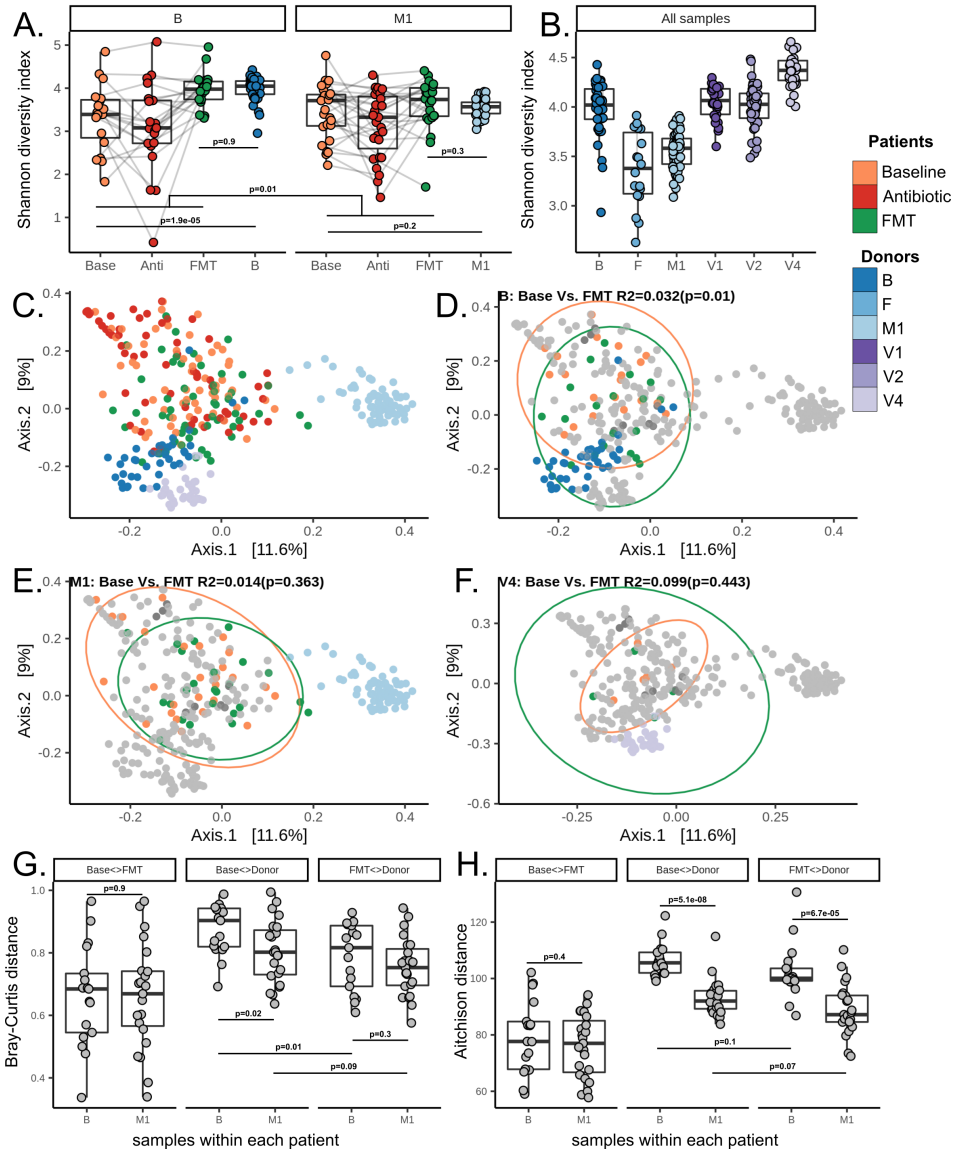

FIGURE 5.5: Comparison of donor B versus M1 in inducing microbial change. **A.** Comparison of the Shannon index for samples collected from patients baseline (Base), post-antibiotc (Anti), post-FMT (FMT), and donors. The left and right facets shows patients who received donor B and donor M1 FMT slurries. **B.** Comparison of the Shannon diversity across donor samples. PCoA of Bray-Curtis, beta diversity, distances between all samples (**C.**), baseline and post-FMT for patient who received FMT from donor B (**D.**), M1 (**E.**), and V4 (**F.**)

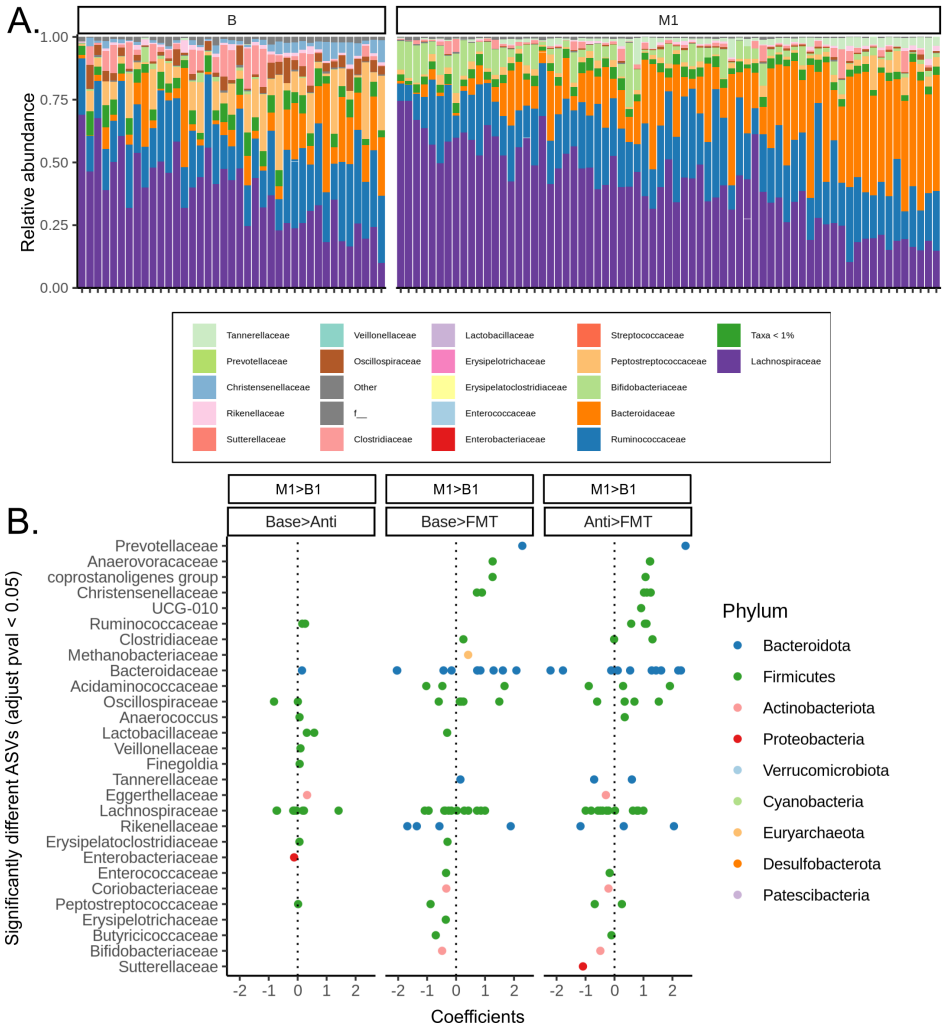

FIGURE 5.6: Taxonomic composition of fecal slurries collected from donors B and M1. **A.** Relative abundance of bacterial families. **B.** Significantly different ASVs between baseline and post-FMT (FMT) samples collected from patients who received donor B and M1 slurries are shown for each family in y-axis. The families are ordered based on mean estimates of differences. x-axis shows a natural log of coefficients.

5.3.4 Microbial engraftment post-FMT: donor B vs. M1

In order to examine whether the observed microbial changes in donor B patients are the result of microbial engraftment, donor ASVs that were engrafted post-FMT from the two major donors (B and M1) were identified. To do this, ASVs that were specific to each donor were compared with data from pre- and post-FMT in patients. In this model, donor ASV with relative abundance of 0 pre-FMT and $> 0.01\%$ post-FMT were defined as engrafted. "Engrafted" ASVs from donors B and M1 in patients with different donors, was used to estimate the rate of spurious "engraftment". The engrafted ASVs were visualized in an increasing number of patients to find whether a group of these ASVs were commonly engrafted despite the microbial variation in each patient (Fig. 5.7A). Since donors B and M1 were used for different numbers of patients (B: 19, M1: 24), 15 patients from each donor were randomly subsampled 100 times with replacement and re-calculated engrafted ASVs (Fig. 5.7B). Our results showed that the engrafted ASVs for both B and M1 donors contains spurious engraftment — ASVs that were detected from wrong donor post-FMT — in which the donor's effect is less clear. However, donor B's ASVs were observed in an increased number of engraftments compared to donor M1 (Fig. 5.7B).

5.4 Discussion

Current therapies for UC patients are primarily focused on suppressing the immune response without targeting the main trigger of the inflammation (Talley et al. 2011). The etiology of UC is complex, but the intestinal microbiome is the environmental factor most closely related to UC. Previously, we have shown the efficacy of FMT in a RCT to induce remission in active UC patients (24% FMT vs. 5% placebo (Moayyedi et al. 2015)). Systematic reviews of previous RCTs have shown that antibiotic therapy has a potential effect on reducing disease activity in UC (Khan et al. 2011). However, as different antibiotics were used in each trial, it is difficult to understand whether

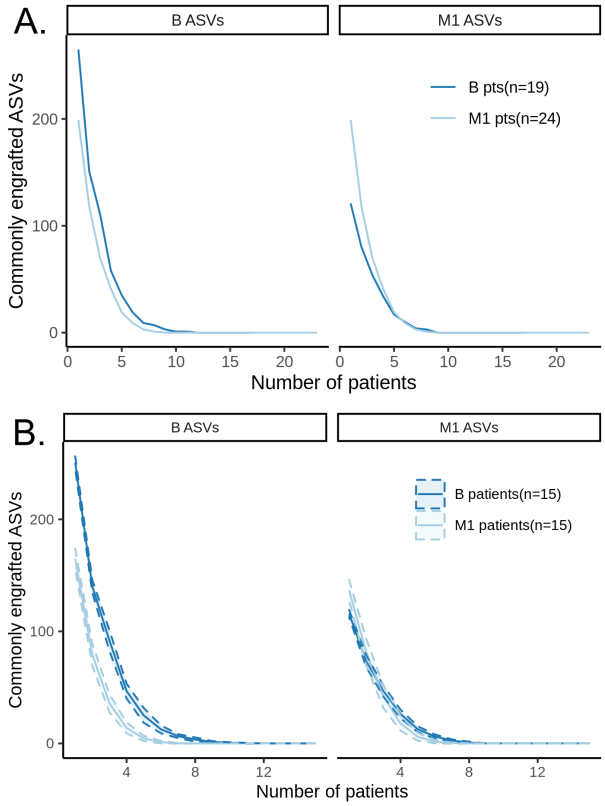

FIGURE 5.7: Effects of donor on microbial engraftment post-FMT. **A.** Donor B and M1 ASVs that were commonly engrafted across patients who received donor B and M1 slurries in an increasing number of individuals post-FMT. **B.** Donor B and M1 ASVs that were commonly engrafted across 15 patients (100 subsampling) who received donor B and M1 slurries.

suppressing a group of bacteria induces more remission in UC patients. In this study, the RCT was designed to determine whether pretreatment with antibiotics would increase the efficacy of FMT. Two mechanistic rationales for antibiotic pretreatment are 1) that the antibiotics may reduce or eliminate pathogenic bacteria that contribute to disease, 2) depletion of gut microbiota by the antibiotics might improve engraftment of donor microbiota. While the final clinical report on this RCT are still pending, there does not appear to be a benefit to the course of antibiotic prior to FMT.

We found that antibiotic therapy significantly reduced microbial diversity and

changed the microbial composition of patients. We observed that the microbial composition of patients who received either placebo or antibiotic pretreatment changed after FMT, suggesting that FMT had a more substantial effect than antibiotic therapy. Although studies have suggested a role of pathogenic bacteria in UC (Petersen et al. 2009; Mirsepasi-Lauridsen et al. 2016; Axelrad et al. 2018), it is still not clear what those bacteria are and how they are involved in mucosal inflammation. We used two broad spectrum antibacterial compounds with some activity against parasites and an antifungal (metronidazole and doxycycline, and terbinafine, respectively). We could not assess mucosal appearance at the end of trial for seven patients and they were excluded from our analysis. Nevertheless, our results showed that 33% and 28% of patients who received only FMT therapy and antibiotic pretreatment before FMT, respectively went to remission suggesting that antibiotics do not improve FMT in the treatment of UC.

16S rRNA gene amplicon sequencing has been used to detect microbial engraftment post-FMT (Khanna et al. 2017b; Hamilton et al. 2013; Staley et al. 2019; Staley et al. 2021). We investigated whether this approach provides adequate resolution to examine microbial engraftment by tracking donor-specific ASVs in FMT recipients. We used data from two donors (B and M1) who provided FMT to the highest number of patients (B:19, M1:24). It was shown that the microbial changes post-FMT were individual-specific. To address these variations, we assessed engrafted ASVs in an increasing number of patients (common engraftment). We found that the difference between engraftment (matched donor) and spurious engraftment (non-matched donor) was not distinguishable, indicating that 16S rRNA gene sequencing does not provide enough resolution to detect engraftments. As I report in Chapter 3 and inconsistent with previous metagenomic studies (Smillie et al. 2018; Paramsothy et al. 2019; Chu et al. 2021; Podlesny et al. 2022b), FMT induces strain-level microbial changes, and high-resolution microbiome analysis is required to detect these subtle changes. More recently, It was implicated that

antibiotic pretreatment in rCDI patients reduces colonization resistance and leads to increased microbial engraftment (Podlesny et al. 2022b). We have conducted metagenomic seqeuncing for all the patients and donors involved in this RCT. We will use this dataset to examine whether antibiotic pretretmanet is associated with microbial engraftment in UC.

In our trial, we had two major donors (donors B and M1) who provided FMT to the highest number of patients (B: 19, M1: 24). Our results showed that donor B was more successful than M1 in shifting microbial composition and the patients' microbiomes became more similar to donor B post-FMT. Although we observed a high rate of suprious engraftment in detecting commonly engrafted ASVs, the number of donor B's engrafted ASVs was greater than M1. Our results, consistent with previous findings (Moayyedi et al. 2015; Wilson et al. 2021), suggests that the donor microbiome affects microbial changes post-FMT. However, we argue that the extent of this effect needs to be investigated with a higher-resolution metagenomic analysis.

Chapter 6

Conclusions

Within the work of this thesis, I present in-depth gut microbiome characterization through culture-independent and -dependent sequencing of healthy individuals and patients with ulcerative colitis (UC). The focus of this thesis was to develop and improve computational approaches to study intestinal microbiota with the goal of shifting from microbiome associations to causation in human health and disease. In **Chapter 2**, I developed a bioinformatics workflow to apply shotgun metagenomics to comprehensive culture-enrichment of the intestinal microbiota and compared this approach to culture-independent (direct) metagenomics from the same samples. I show that culture-enriched metagenomics (CEMG) improves *de novo* assembly of the gut microbiota compared to direct metagenomics (DMG) by providing a more in-depth view of microbial genes and genomes using data from eight healthy individuals.

In **Chapter 3**, I applied CEMG to a successful fecal microbiota transplantation (FMT) donor based on a randomized controlled trial (RCT) for UC patients (Moayyedi et al. 2015). The higher resolution provided by CEMG allowed us to identify a group of genes commonly engrafted in patients who responded to FMT. Using publicly available genomes and metagenomic datasets, I show that most of these genes were strain-specific and over-represented in the healthy individuals than UC patients **(Chapter 3)**.

Tracking non-bacterial component of microbiota, such as bacteriophages, is essential and can affect FMT outcomes. In **Chapter 4**, I present a highly dynamic bacteriophage, crAssphage, using longitudinal data from an FMT donor. I developed a pipeline to track the crAssphage strain present in donors based on SNP information, and I show that accurate detection of bacteriophage engraftment post-FMT requires SNP analysis in UC that PCR detection is not sufficient evidence for engraftment. And in **Chapter 5**, we report the first RCT to assess the efficacy of antibiotic treatment prior to FMT in UC patients. We showed that antibiotic therapy changed the microbial composition but didn't improve the efficacy of FMT.

The previous work from the Surette lab showed that the culture-enrichment provides a more robust assessment of the human lung microbiota (Sibley et al. 2011; Whelan et al. 2020). Although we previously published the same protocol for the molecular profiling of the gut microbiota (Lau et al. 2016), I applied shotgun metagenomics to this approach for the first time. The intestinal microbiota is significantly more diverse than lung microbiota, and the complex dataset generated by this approach required a new bioinformatics pipeline. I have compared the performance of widely used metagenomic algorithms, some of which were not presented in this thesis, but they were instrumental for this body of work. These comparisons included *de novo* assembler, binning algorithms, taxonomic assigner, and functional annotation approaches. *De no* assembly algorithms, particularly de Bruijn methods, that are standard in the field and have advanced the microbial genome collections are highly computationally expensive. I compared multiple assembly approaches, including sub-assembly, co-assembly and single assembly of cultured plate pools, to develop and optimize memory-intensive cloud instances (google cloud) for CEMG assemblies. The results showed that co-assembly of plate pools with metaSPAde produced the highest quality assemblies (measured by N50) with the shortest run time.

Recent *de novo* assembled catalogues of genes (Coelho et al. 2022) and genome (Almeida et al. 2021) from metagenomic samples has advanced the field and provided an extensive resource for hypothesis generation. Nevertheless, these collections may provide a spurious interpretation of the human microbiota. Even the most conservative gene prediction program, such as Prodigal, misses up to 5% of genes and consequently functions. This is exaggerated with highly fragmented contigs, which results in incomplete open reading frames (ORFs) generated by *de novo* assembly with short read sequencing. Further, there is no robust metric to assess the quality of metagenome assembled genomes (MAGs), and they are often incomplete or contain multiple strains in the same bin. A single metagenomic sample likely includes thousands of strains, and the uniform coverage information required by binning algorithms is often missing in these datasets. CEMG provide more complete assembly fragments and unique coverage information from multiple plate pools that can improve *de novo* assembly and binning of contigs.

The widely used prokaryote assembler that use de Bruijn graph-based methods divides a read into k-mer sequences to construct a graph. In a more complex microbial community (e.g. deep sequencing of gut microbiota or environmental samples), the split k-mers might result in misassemblies. Although debatable and not fully understood, there seems to be a threshold where the increased depth of sequencing (or co-assembly) results in a highly complex assembly graph which causes increased misassemblies in the contigs. The third-generation high-throughput sequencing (HTS) methods (such as Nanopore and PacBIO) could address some of these challenges and we should focus on combining this method with CEMG in future. Particularly, assembling these long read sequences via string-based algorithms that avoid dividing reads will potentially result in a greater quality of genes and genomes. The long-read sequencing is more expensive and less standardized. It was shown that optimizing the genome library preparation could reduce the cost of these methods (Derakhshani et al. 2020) and, in future, should be followed for metagenomics library construction of PacBIO sequencer.

It is evident that studying microbial changes post-FMT, particularly determining the donor's effect on recipients' microbiota, requires strain-level resolution (Li et al. 2016; Smillie et al. 2018; Podlesny et al. 2022b). However, the term "strain" is not well defined, and it is highly debatable to describe a standard genomic data property that most accurately represents a strain from a microbial community. In classical microbiology, strain is an isolate from pure culture that originated from a single colony (Dijkshoorn et al. 2000); however, this definition is more flexible in microbial genetics which defined by phylogenetic principles originally derived from eukaryote taxonomy (Hugenholtz et al. 2021). Typically it is expected to observe 95-97% identity in core genes of a species while the identity threshold could increase to >99-99.9% to be considered a strain. Similar to the terminology, the methods attempted to identify strains in FMT studies were debatable and not standardized.

Although 16S rRNA gene amplicon sequencing does not provide strain-level resolution, it has been the most widely used approach to track microbial changes post-FMT. Even the full 16S rRNA gene can not distinguish closely related species or strains from each other. For example, *Escherichia coli* (*E. coli*) and *Shigella sp.* have an almost identical 16S rRNA gene (Brenner et al. 1972; Ragupathi et al. 2018). Using data from two RCTs, I show that 16S rRNA sequencing is not sufficient to detect donor-specific amplicon sequence variants (ASVs) in FMT recipients **(Chapter 3 and 5)**. The so-called "engrafted ASVs" were detected independent of the recipient-specific microbial changes, which were determined by common engraftment across an increasing number of patients. I compared these expected engraftments from a matched donor to a placebo treatment **(Chapter 3)** or a non-matched donor **(Chapter 5)**. These studies provide an approach to measure noise in FMT experiments and indicate that 16S rRNA sequencing does not provide sufficient resolution to determine donor-specific ASVs. The patient-specific microbial changes post-FMT were evident by 16S rRNA sequencing. However, shotgun metagenomics was required to track microbial changes and investigate microbial

engraftment post-FMT.

Marked-based approaches are predominately used to detect microbial strains in metagenomic data from FMT studies. The tools such as StrainPhlan (Truong et al. 2017), strainFinder (Smillie et al. 2018) , PStrain (Wang et al. 2021), and SameStr (Podlesny et al. 2022a) rely on marker gene databases (e.g., MetaPhlAn) to identify species-level markers and use SNV information by read mapping to infer strains. These methods are sensitive to the sequencing depth and limited to the most abundant strain within each species. Alternatively, kmer-based approaches such as GT-Pro (Shi et al. 2022), and StrainGE (Dijk et al. 2022) have been developed that work based on unique kmer information. Although these methods are computationally efficient, there seems to be a trade-off in the length of k-mer in which longer n provides higher resolution but with the cost of reduced sensitivity. Even if we assume maker-based approaches determine 100% of strains in a metagenomic sample, like multilocus sequence typing (MLST), they provide no functional information about the so-called strains in FMT studies and gene content and phenotype can vary within a single MLST sequence type.

Additionally, assembly-based approaches have been used to detect microbial strains. It is debatable whether MAGs represents strains, but MAGs of the same species from different sources probably represent different strains. For each assembly, contigs and MAGs represent consensus assemblies. While SNP analysis can be used to estimate strains diversity based on core genes, accessory genes (which define functional differences between strains) tend to be excluded from MAGs if multiple strains are present in a sample. CEMG improves this over DMG as I show in **Chapter 2** with the increase size of MAGs in the CEMG assemblies. However, these are still smaller than genomes from isolated strains. High-quality MAGs resolved by CEMG were used to track strains post-FMT **(Chapter 3)**. Similar to the past studies (Lee et al. 2017; Watson et al.

2021), short reads from recipients were mapped to the MAGs collection and 1X coverage information cut-offs were used to detect engrafted, and replaced MAGs post-FMT **(Chapter 3)**. Importantly, this approach provides functional context for the strains but will be limited to the dominant strain in the community based on the consensus assemblies. We show that CEMG can potentially address some of these challenges **(Chapter 3)** by binning low abundant microbes. The other caveat is that the short reads can map to multiple MAGs in the reference. Potentially, this issue can be resolved by increasing mapping stringency, filtering primarily perfectly mapped reads. More recently, algorithms such as STRONG (Quince et al. 2021) and SynTracker (Enav and Ley 2021) have attempted to address some of these challenges by strain decomposition of only core-genes and pairwise comparison of homologous regions, respectively. However, there is no consensus on defining single-copy core genes for each strain, and accessory genes are disregarded in these approaches.

As mentioned above, many tools have attempted to track microbial strains, particularly after FMT treatment, but less effort was made to validate these approaches. In future, synthetic mock metagenomes from single whole-genomic data with variable degrees of microbial complexity should be generated to investigate whether these approaches are adequately robust. Further, culture-enriched plate pools from selective media can be used to test the recovery of closely-related strains. The golden standard to evaluate the efficacy of a treatment in medicine is to conduct RCT. Similarly, metagenomic data from patients who received a placebo should be used to estimate the error rate and validate strain engraftment post-FMT. I think the Surette lab is well positioned to follow these projects in future. FMT is not risk-free, and the field should move from FMT therapy to small molecule therapies or defined communities based on FMT results. To do this, investigating the functional mechanism of strain colonization should be followed instead of only methods that track strains in FMT studies.

With improvement in our publicly available sequence repositories, it has become easier to re-analyze and merge various omics datasets from FMT studies (e.g. recurrent-*Clostridioides difficile* infection (rCDI), metabolic disorder, UC, Crohn's disease (CD)). However, these diseases manifest different phenotype which suggests the mechanism of action and possibly the importance of the donor depends on the disease. For example, CDI is an acute infectious disease, while IBD is a chronic inflammatory disease. In rCDI patients, the goal of FMT is to restore the microbial community balance, and it was shown that independent of the donor's microbial composition, the recipient's microbiome becomes more similar to the donor post-FMT, although evidence for engraftment is often weak. In other GI-related diseases such as UC, we have more heterogeneity in the microbial composition of recipients. The patient's microbiome likely changes post-FMT but these differences are not as stark as rCDI. As a result, the mechanism of FMT seems to be more complex. For example, Podlesny et al. 2022b recently suggested that antibiotic pretreatment leads to increased strain engraftment post-FMT using multiple omics datasets from rCDI patients. In contrast, we presented that the antibiotic pretreatment in UC does not affect FMT outcomes but significantly changes the patient's microbiota in a RCT **(Chapter 5)**. In future, metagenomic sequencing should be carried out to investigate whether there is any association between microbial colonization and antibiotic pretreatment in UC. However, given the concerns with antibiotic resistance (Chatterjee et al. 2018; Laxminarayan et al. 2020), antibiotic treatment before FMT should be carefully recommended only based on disease manifestation, instead of a standard protocol to increase the efficacy of FMT. Another example is the bacteriophage colonization post-FMT. It was implicated that crAssphage is a stable phage that colonizes rCDI patients post-FMT (Draper et al. 2018; Siranosian et al. 2020). However, I presented a dynamic crAssphage in a healthy donor that was not engrafted in UC patients post-FMT with data from a subset of FMT participants **(Chapter 4)** indicating that potentially bacteriophage colonization after FMT is related to disease phenotypes.

The intestinal microbiome is an essential part of our health. This thesis further provides insight into these microbial communities, their functions, and balance in healthy individuals and UC patients. In past decades, an enormous number of studies have characterized the gut microbiota using culture-independent approaches. The work of this thesis shows that classical microbiology, in combination with metagenomics, provides an opportunity to improve our informatics methods to characterize gut microbiota. The role of intestinal microbiota in UC patients is evident, and FMT has emerged as a potential therapy for these patients. The data and results presented within suggest that a high-resolution microbiome analysis is required to understand the mechanism of bacterial and non-bacterial colonization post-FMT. Ultimately, FMT is not an appealing treatment for patients, and the field should transition to new microbial-based therapies. Culture-enriched metagenomic coupled with new sequencing technologies can truly help us to do this transition and better understand the mechanism of action post-FMT.

Appendix A

Chapter 2 Supplement

TABLE A1.1: List of culture-enriched plates and stool samples selected for metagenomic sequencing.

Number	Sample	Donor	Class	Media	Paired-end	Source
1	S1C1	SHCM1	anaerobic	AIAana	12423664	Plate pool
2	S1C2	SHCM1	anaerobic	BHI4ana	6972664	Plate pool
3	S1C3	SHCM1	aerobic	BHI5aer	6540709	Plate pool
4	S1C4	SHCM1	anaerobic	BHI5ana	6074129	Plate pool
5	S1C5	SHCM1	aerobic	CHOCaer	6065023	Plate pool
6	S1C6	SHCM1	anaerobic	FAAana	6723345	Plate pool
7	S1C7	SHCM1	anaerobic	KVLBana	9958984	Plate pool
8	S1C8	SHCM1	anaerobic	M9inuana	11817414	Plate pool
9	S1C9	SHCM1	anaerobic	M9mucana	13908034	Plate pool
10	S1C10	SHCM1	anaerobic	M9pectana	10962936	Plate pool
11	S1C11	SHCM1	aerobic	Mkaer	6016412	Plate pool
12	S1C12	SHCM1	anaerobic	Mkana	6443660	Plate pool
13	S1C13	SHCM1	anaerobic	MRSana	6618148	Plate pool
14	S1C14	SHCM1	aerobic	PEAaer	9599360	Plate pool
15	S2C1	SHCM2	aerobic	ppae1012	10391787	Plate pool
16	S2C2	SHCM2	aerobic	ppae1427	14648038	Plate pool
17	S2C3	SHCM2	aerobic	ppae16	12934202	Plate pool
18	S2C4	SHCM2	aerobic	ppae17	11593021	Plate pool
19	S2C5	SHCM2	aerobic	ppae30	11595422	Plate pool
20	S2C6	SHCM2	aerobic	ppae3233	15251439	Plate pool
21	S2C7	SHCM2	anaerobic	ppana10	16348397	Plate pool
22	S2C8	SHCM2	anaerobic	ppana13	13981690	Plate pool
23	S2C9	SHCM2	anaerobic	ppana16	16259303	Plate pool
24	S2C10	SHCM2	anaerobic	ppana17	15436604	Plate pool
25	S2C11	SHCM2	anaerobic	ppana18	12703011	Plate pool
26	S2C12	SHCM2	anaerobic	ppana20	10090848	Plate pool
27	S2C13	SHCM2	anaerobic	ppana22	10630986	Plate pool
28	S2C14	SHCM2	anaerobic	ppana25	6084039	Plate pool
29	S2C15	SHCM2	anaerobic	ppana26	14549152	Plate pool
30	S2C16	SHCM2	anaerobic	ppana29	23417489	Plate pool
31	S2C17	SHCM2	anaerobic	ppana2	13161099	Plate pool
32	S2C18	SHCM2	anaerobic	ppana31	15928472	Plate pool
33	S2C19	SHCM2	anaerobic	ppana3	20094719	Plate pool
34	S2C20	SHCM2	anaerobic	ppana6	14803557	Plate pool

Number	Sample	Donor	Class	Media	Paired-end	Source
35	S3C1	SHCM3	anaerobic	AIAana	15079290	Plate pool
36	S3C2	SHCM3	anaerobic	BBEana	11470669	Plate pool
37	S3C3	SHCM3	anaerobic	BEEFana	8710803	Plate pool
38	S3C4	SHCM3	anaerobic	BHI2ana	9332890	Plate pool
39	S3C5	SHCM3	anaerobic	BHICELLana	11413552	Plate pool
40	S3C6	SHCM3	anaerobic	BSMana	10262653	Plate pool
41	S3C7	SHCM3	anaerobic	GIFUana	9593351	Plate pool
42	S3C8	SHCM3	anaerobic	GMMana	7152566	Plate pool
43	S3C9	SHCM3	anaerobic	KVLBana	8023341	Plate pool
44	S3C10	SHCM3	anaerobic	M9INUana	9742779	Plate pool
45	S3C11	SHCM3	anaerobic	MKana	7884008	Plate pool
46	S3C12	SHCM3	anaerobic	MRSana	7962911	Plate pool
47	S3C13	SHCM3	aerobic	MRSar	13571107	Plate pool
48	S3C14	SHCM3	anaerobic	MSAana	8985249	Plate pool
49	S4C1	SHCM4	anaerobic	AIAana	8560850	Plate pool
50	S4C2	SHCM4	anaerobic	BHI3ana	11617084	Plate pool
51	S4C3	SHCM4	anaerobic	BHI5ana	10857541	Plate pool
52	S4C4	SHCM4	anaerobic	BHICELLana	10360435	Plate pool
53	S4C5	SHCM4	anaerobic	BHIINUana	10904630	Plate pool
54	S4C6	SHCM4	anaerobic	BHIMUCana	8688033	Plate pool
55	S4C7	SHCM4	anaerobic	BHIPECana	11712255	Plate pool
56	S4C8	SHCM4	anaerobic	BSMana	9623876	Plate pool
57	S4C9	SHCM4	anaerobic	CANana	8764280	Plate pool
58	S4C10	SHCM4	anaerobic	GIFUana	10040569	Plate pool
59	S4C11	SHCM4	anaerobic	GMMana	11192751	Plate pool
60	S4C12	SHCM4	anaerobic	M9INUana	12191238	Plate pool
61	S4C13	SHCM4	anaerobic	M9MUCana	11952205	Plate pool
62	S4C14	SHCM4	anaerobic	MRSana	10315073	Plate pool
63	S4C15	SHCM4	anaerobic	TSYana	11465462	Plate pool
64	S5C1	SHCM5	anaerobic	BBEana	8434003	Plate pool
65	S5C2	SHCM5	anaerobic	BHI2ana	10985548	Plate pool
66	S5C3	SHCM5	anaerobic	BHI5ana	9776659	Plate pool
67	S5C4	SHCM5	anaerobic	BHIMUCana	10578043	Plate pool
68	S5C5	SHCM5	anaerobic	CBAana	8477519	Plate pool
69	S5C6	SHCM5	aerobic	CNAaer	7019320	Plate pool
70	S5C7	SHCM5	anaerobic	FAAana	8318595	Plate pool

Number	Sample	Donor	Class	Media	Paired-end	Source
71	S5C8	SHCM5	anaerobic	GIFUana	8464269	Plate pool
72	S5C9	SHCM5	aerobic	GMMaer	8777563	Plate pool
73	S5C10	SHCM5	anaerobic	KVLBana	8747468	Plate pool
74	S5C11	SHCM5	anaerobic	M9PECTana	7965356	Plate pool
75	S5C12	SHCM5	anaerobic	MKana	13338529	Plate pool
76	S5C13	SHCM5	aerobic	MSAaer	8469224	Plate pool
77	S5C14	SHCM5	anaerobic	TSYana	8681989	Plate pool
78	S6C1	SHCM6	anaerobic	BBEana	8813192	Plate pool
79	S6C2	SHCM6	anaerobic	BEEFana	7863172	Plate pool
80	S6C3	SHCM6	anaerobic	BHI3ana	11550569	Plate pool
81	S6C4	SHCM6	anaerobic	BHI5ana	11122559	Plate pool
82	S6C5	SHCM6	anaerobic	BHICellana	7902431	Plate pool
83	S6C6	SHCM6	anaerobic	CBAana	8704711	Plate pool
84	S6C7	SHCM6	aerobic	CHOCaer	8125876	Plate pool
85	S6C8	SHCM6	anaerobic	CHOCana	12422595	Plate pool
86	S6C9	SHCM6	anaerobic	CNAana	9443836	Plate pool
87	S6C10	SHCM6	anaerobic	GMMana	4222141	Plate pool
88	S6C11	SHCM6	anaerobic	KVLBana	11422493	Plate pool
89	S6C12	SHCM6	anaerobic	MACana	8987424	Plate pool
90	S6C13	SHCM6	aerobic	MixedAer	10515677	Plate pool
91	S6C14	SHCM6	anaerobic	MKana	10911723	Plate pool
92	S6C15	SHCM6	anaerobic	MRSana	14659766	Plate pool
93	S6C16	SHCM6	anaerobic	MSAana	15347614	Plate pool
94	S15C1	SHCM15	aerobic	aer30	15333183	Plate pool
95	S15C2	SHCM15	aerobic	aer4	23667949	Plate pool
96	S15C3	SHCM15	anaerobic	ana10	22442273	Plate pool
97	S15C4	SHCM15	anaerobic	ana10b	166163330	Plate pool
98	S15C5	SHCM15	anaerobic	ana11	16142855	Plate pool
99	S15C6	SHCM15	anaerobic	ana12	30020329	Plate pool
100	S15C7	SHCM15	anaerobic	ana15	66365295	Plate pool
101	S15C8	SHCM15	anaerobic	ana16	18831075	Plate pool
102	S15C9	SHCM15	anaerobic	ana18	40436742	Plate pool
103	S15C10	SHCM15	anaerobic	ana20	35197675	Plate pool
104	S15C11	SHCM15	anaerobic	ana23	67231962	Plate pool
105	S15C12	SHCM15	anaerobic	ana24	43118477	Plate pool

Number	Sample	Donor	Class	Media	Paired-end	Source
106	S15C13	SHCM15	anaerobic	ana26	22041105	Plate pool
107	S15C14	SHCM15	anaerobic	ana28	22275048	Plate pool
108	S15C15	SHCM15	anaerobic	ana29	11189647	Plate pool
109	S15C16	SHCM15	anaerobic	ana31	26292516	Plate pool
110	S15C17	SHCM15	anaerobic	ana7	47788248	Plate pool
111	S15C18	SHCM15	anaerobic	ana8	15328894	Plate pool
112	S15C19	SHCM15	anaerobic	ana9	26285894	Plate pool
113	B13C1	SHCM0	Other	P10	7212641	Plate pool
114	B13C2	SHCM0	Other	P11	13658127	Plate pool
115	B13C3	SHCM0	Other	P1	13547295	Plate pool
116	B13C4	SHCM0	Other	P12	9955109	Plate pool
117	B13C5	SHCM0	Other	P13	12746585	Plate pool
118	B13C6	SHCM0	Other	P2	8114888	Plate pool
119	B13C7	SHCM0	Other	P3	15246685	Plate pool
120	B13C8	SHCM0	Other	P4	14464941	Plate pool
121	B13C9	SHCM0	Other	P5	18023885	Plate pool
122	B13C10	SHCM0	Other	P6	19748129	Plate pool
123	B13C11	SHCM0	Other	P7	19266475	Plate pool
124	B13C12	SHCM0	Other	P8	12827228	Plate pool
125	B13C13	SHCM0	Other	P9	12450314	Plate pool
126	S1S1	SHCM1			18893172	Stool
127	S2S1	SHCM2			29250143	Stool
128	S3S1	SHCM3			43115009	Stool
129	S4S1	SHCM4			28836989	Stool
130	S5S1	SHCM5			28774192	Stool
131	S6S1	SHCM6			40326457	Stool
132	S15S1	SHCM15			116995473	Stool
133	B13	SHCM0			19890046	Stool
134	B16	SHCM0			7629264	Stool

Appendix B

Chapter 3 Supplement

TABLE A2.1: List of characterized proteins from commonly engrafted genes.

#	Protein.names	Pfam	Taxonomic Family
1	Single-stranded DNA-binding protein	PF00436;	Lachnospiraceae
2	AAA domain-containing protein		Lachnospiraceae
3	ESAT-6-like protein	PF06013;	Lachnospiraceae
4	ESAT-6-like protein	PF06013;	Lachnospiraceae
5	(4Fe-4S)-binding protein	PF00037;PF01243;	Lachnospiraceae
6	Signal peptidase I W (EC 3.4.21.89)	PF00717;	Lachnospiraceae
7	LPD11 domain-containing protein	PF18824;	Lachnospiraceae
8	C2H2-type domain-containing protein		Lachnospiraceae
9	Type IV pilus twitching motility protein PilT	PF00437;	Lachnospiraceae
10	D-ribose-binding periplasmic protein	PF13407;	Clostridiaceae
11	Ribosomal protein HS6-type (S12/L30/L7a)	PF01248;	Clostridiaceae
12	Integral membrane protein (Intg_mem_TP0381)	PF09529;	Clostridiaceae
13	Probable membrane transporter protein	PF01925;	Clostridiaceae
14	Sodium/proline symporter (Proline permease)	PF00474;	Clostridiaceae
15	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	PF00528;PF12911;	Clostridiaceae
16	Galactoside transport system permease protein mglC	PF02653;	Clostridiaceae
17	4HBT domain-containing protein	PF03061;	Clostridiaceae
18	Probable cell division protein ytgP	PF01943;	Lachnospiraceae
19	ANTAR domain protein	PF03861;	Lachnospiraceae
20	Sugar-specific transcriptional regulator, TrmB family	PF01978;	Lachnospiraceae
21	Pyridoxine kinase (EC 2.7.1.35)	PF08543;	Lachnospiraceae
22	Enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (2-phosphoglycerate dehydratase)	PF00113;PF03952;	Lachnospiraceae
23	Flp pilus assembly protein, protease CpaA (Prepilin peptidase)	PF01478;	Lachnospiraceae
24	Stage V sporulation protein AB	PF13782;	Lachnospiraceae
25	Biotin transporter	PF02632;	Lachnospiraceae
26	Chorismate–pyruvate lyase		Lachnospiraceae
27	Cold shock domain-containing protein (Cold shock-like protein)	PF00313;	Lachnospiraceae
28	SseB domain-containing protein	PF07179;	Lachnospiraceae
29	Cell division inhibitor MinD	PF13614;	Lachnospiraceae
30	Cell division suppressor protein YneA		Lachnospiraceae
31	RNA polymerase sigma factor sigX	PF04542;PF08281;	Lachnospiraceae
32	HIT-like protein (EC 3.-.-.-)	PF01230;	Lachnospiraceae
33	Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23)	PF00692;	Lachnospiraceae
34	Uncharacterized conserved protein	PF12821;	Lachnospiraceae
35	Sporulation protein, YlmC/YmxH family (YlmC/YmxH family sporulation protein)	PF05239;	Lachnospiraceae
36	tRNA-dihydrouridine synthase (EC 1.3.1.-)	PF01207;	Lachnospiraceae
37	RNA polymerase sigma factor	PF04542;PF04545;	Lachnospiraceae
38	Phosphatidylglycerol lysyltransferase (EC 2.3.2.3) (Lysylphosphatidylglycerol synthase)	PF03706;	Lachnospiraceae
39	Molybdate ABC transporter, permease protein	PF00005;	Lachnospiraceae
40	Septum formation initiator (Septum formation initiator family protein)	PF04977;	Lachnospiraceae
41	N-acetylmuramoyl-L-alanine amidase LytC (EC 3.5.1.28)	PF01520;	Lachnospiraceae
42	tRNA-specific adenosine deaminase (EC 3.5.4.33)	PF14437;	Lachnospiraceae
43	Dicarboxylate/amino acid:cation symporter (Glutamate-aspartate carrier protein)	PF00375;	Lachnospiraceae
44	Flagellin N-methylase	PF03692;	Lachnospiraceae
45	Cytidylate kinase		Lachnospiraceae
46	Ubiquitin-like domain-containing protein	PF03780;	Lachnospiraceae
47	Formate channel 1 (Formate/nitrite transporter family protein)	PF01226;	Lachnospiraceae
48	2-dehydro-3-deoxy-6-phosphogalactonate aldolase (EC 4.1.2.21)	PF01081;	Lachnospiraceae
49	Cytidylate kinase		Lachnospiraceae
50	Hpt domain	PF01627;	Lachnospiraceae
51	[Ribosomal protein S18]-alanine N-acetyltransferase (EC 2.3.1.266)	PF00583;	Lachnospiraceae
52	DUF1275 domain-containing protein	PF06912;	Lachnospiraceae
53	DnaB_2 domain-containing protein	PF07261;	Ruminococcaceae
54	Mini-ribonuclease 3 (Mini-3) (Mini-RNase 3) (EC 3.1.26.-) (Mini-RNase III) (Mini-III)	PF00636;	Ruminococcaceae
55	AbrB family transcriptional regulator (AbrB/MazE/SpoVT family DNA-binding domain-containing protein)	PF04014;	Ruminococcaceae
56	Adenosylcobinamide kinase/adenosylcobinamide-phosphate guanylyltransferase	PF02283;	Ruminococcaceae
57	PTS ascorbate transporter subunit IIC		Ruminococcaceae
58	ANTAR domain-containing protein (Probable transcriptional regulatory protein pdtaR)	PF03861;	Ruminococcaceae
59	DUF5626 domain-containing protein	PF18540;	Ruminococcaceae
60	NADH dehydrogenase (EC 1.6.99.3)	PF00881;	Ruminococcaceae

#	Protein.names	Pfam	Taxonomic Family
61	Putative endoribonuclease L-PSP	PF01042;	Ruminococcaceae
62	DUF2500 domain-containing protein	PF10694;	Ruminococcaceae
63	UPF0145 protein FPR_16870	PF01906;	Ruminococcaceae
64	Na+-driven multidrug efflux pump	PF01554;	Ruminococcaceae
65	Stress-response A/B barrel domain-containing protein	PF07876;	Ruminococcaceae
66	REP element-mobilizing transposase RayT	PF01797;	Ruminococcaceae
67	Oxaloacetate decarboxylase		Ruminococcaceae
68	TrbC/VIRB2 family protein	PF04956;	Ruminococcaceae
69	Nitrous oxide-stimulated promoter	PF11756;	Ruminococcaceae
70	Beta-galactosidase	PF16355;PF18565;PF00703;PF02836;PF02837;	Ruminococcaceae
71	Iron-sulfur cluster carrier protein	PF10609;	Ruminococcaceae
72	Ion channel (Two pore domain potassium channel family protein)	PF07885;	NA
73	Spo0A_C domain-containing protein	PF08769;	NA
74	Predicted transcriptional regulator	PF13443;	NA
75	Cyclic lactone autoinducer peptide		NA
76	Cobyrinic acid a,c-diamide synthase	PF01656;PF07685;	NA
77	DNA-binding helix-turn-helix protein	PF13443;	Lachnospiraceae
78	Signal peptidase I (EC 3.4.21.89)	PF10502;	NA
79	CNA-B domain-containing protein	PF05738;	NA
80	HTH cro/C1-type domain-containing protein	PF13443;	NA
81	Replication initiator A domain-containing protein	PF06970;	NA
82	Pro-sigmaK processing inhibitor BofA	PF07441;	NA
83	L-arabinose transport system permease protein AraQ	PF00528;	Lachnospiraceae
84	Uncharacterized protein		NA
85	Uncharacterized protein		NA
86	Membrane protease subunits stomatin/prohibitin homologs	PF01145;	NA
87	HTH_17 domain-containing protein	PF12728;	NA
88	Translation initiation factor IF-1	PF01176;	NA
89	Cold-shock DNA-binding protein family	PF00313;	NA
90	DUF3991 domain-containing protein	PF13154;	NA
91	Nicotinamide-nucleotide amidohydrolase family protein	PF02464;PF18146;	NA
92	Anti-sigma F factor (EC 2.7.11.1) (Stage II sporulation protein AB)	PF13581;	Ruminococcaceae
93	Urease accessory protein UreF	PF01730;	NA
94	Arginine transport system permease protein ArtQ	PF00528;	NA
95	Exodeoxyribonuclease 7 small subunit (EC 3.1.11.6) (Exonuclease VII small subunit)	PF02609;	NA
96	Bacterial nucleoid DNA-binding protein	PF00216;	NA
97	DUF4230 domain-containing protein	PF14014;	NA
98	Uncharacterized protein		Lachnospiraceae
99	Stage V sporulation protein T	PF04014;PF15714;	NA
100	Phosphoribosylglycinamide formyltransferase		NA
101	Predicted nucleotidyltransferase component of viral defense system	PF08843;	Ruminococcaceae
102	Isopentenyl-diphosphate Delta-isomerase (EC 5.3.3.2)	PF00293;	NA
103	DUF4253 domain-containing protein	PF14062;	Lachnospiraceae
104	RNA polymerase sigma factor sigma-70 family	PF08281;	NA
105	Sensor histidine kinase graS (EC 2.7.13.3)	PF02518;	NA
106	Uncharacterized protein		NA
107	Conserved hypothetical integral membrane protein TIGR02185	PF09605;	NA
108	DUF4367 domain-containing protein	PF14285;	Lachnospiraceae
109	CAI-1 autoinducer sensor kinase/phosphatase CqsS (EC 2.7.13.3)	PF00072;	NA
110	HTH cro/C1-type domain-containing protein	PF01381;	NA
111	IS200/IS605 family transposase	PF01797;	NA
112	50S ribosomal protein L29	PF00831;	Oscillospiraceae
113	Accessory gene regulator protein A	PF04397;PF00072;	Lachnospiraceae
114	GGACT domain-containing protein	PF06094;	Lachnospiraceae
115	Site-specific DNA methylase (EC 2.1.1.37)	PF00145;	Lachnospiraceae
116	50S ribosomal protein L7/L12	PF00542;PF16320;	Lachnospiraceae
117	30S ribosomal protein S10	PF00338;	Lachnospiraceae
118	Bacteriophage Gp15 protein	PF06854;	Ruminococcaceae
119	Putative agmatine deiminase (EC 3.5.3.12) (Agmatine iminohydrolase)	PF04371;	Ruminococcaceae
120	Dinitrogenase iron-molybdenum cofactor	PF02001;PF02579;	Ruminococcaceae
121	HPr domain-containing protein	PF00381;	Ruminococcaceae
122	KipI antagonist	PF02626;	Ruminococcaceae
123	ABC transporter ATP-binding protein	PF00005;PF12399;	Ruminococcaceae
124	4Fe-4S binding domain-containing protein	PF00037;PF12724;	Ruminococcaceae

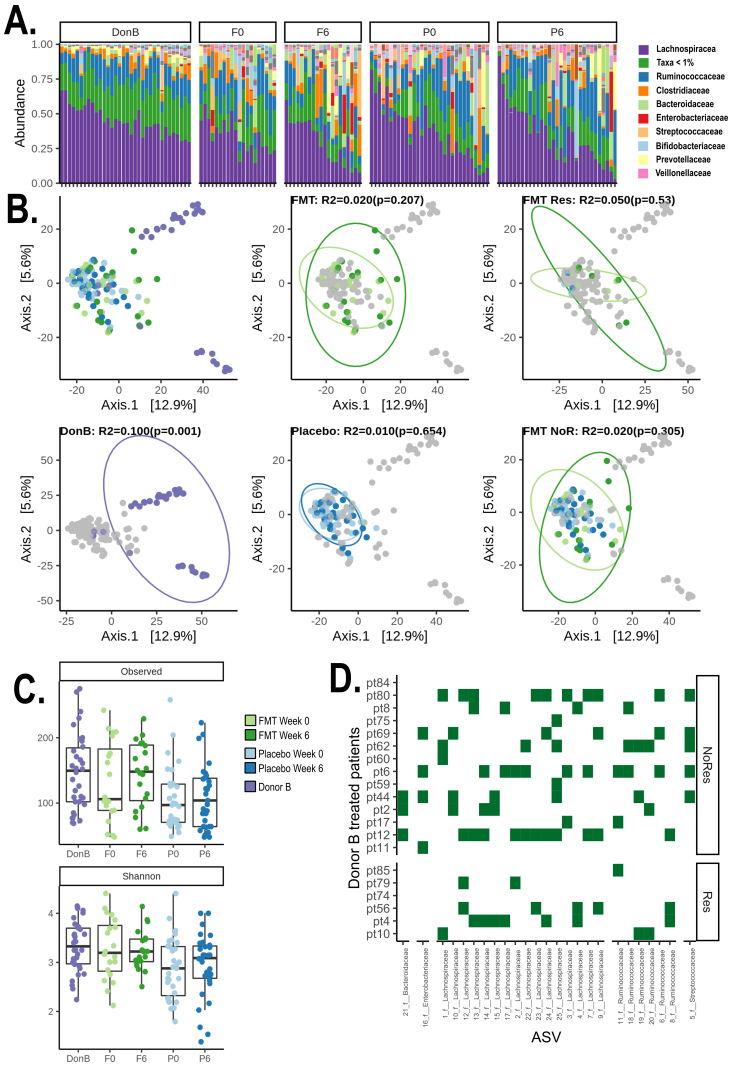

FIGURE A2.1: The microbial composition of 51 patients who either randomly received FMT from donor B or placebo treatment using 16S rRNA gene amplicon sequencing. **A**. Taxonomic composition of samples at the family-level. **B**. PCoA of Aitchison distances for all samples (top left panel), donor B samples compared to all the other samples (bottom left), samples collected from before and after FMT or placebo treatment (middle panels), and samples collected prior/post-FMT in responder (top right) and non-responder (bottom right) patients. **C**. Comparison of observed and Shannon, alpha diversity, metrics for samples collected from patients prior and post FMT or placebo treatment, and donor B. **D**. Comparison of the commonly engrafted ASVs in $\geq$ three individuals across non-responder (NoRes) and responder (Res) patients.

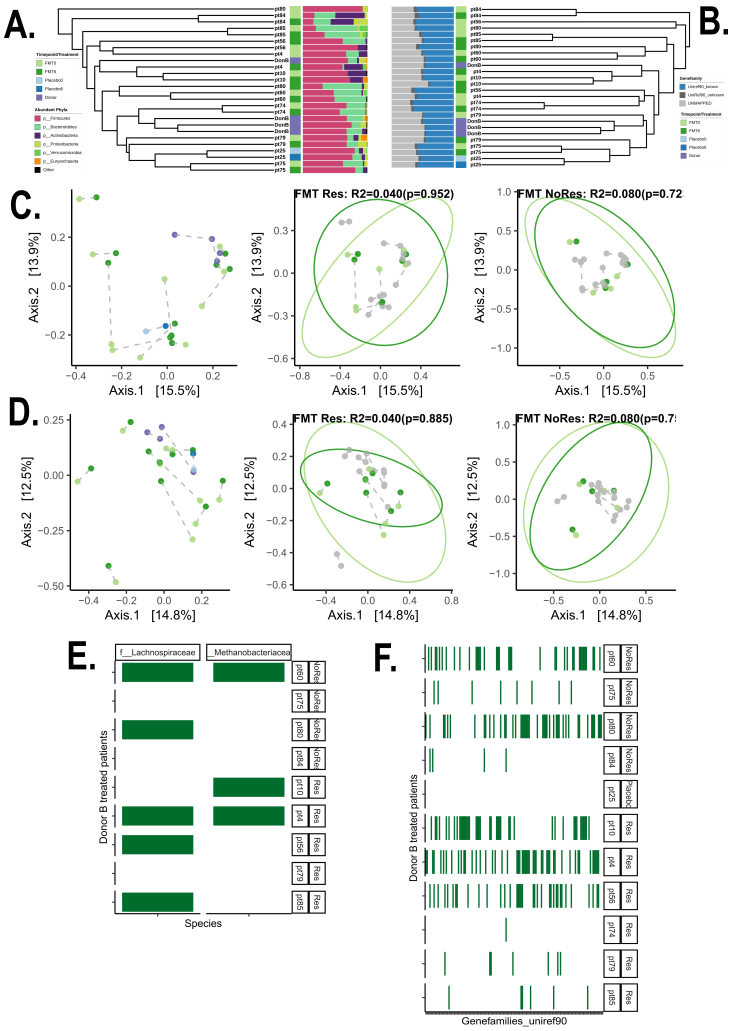

FIGURE A2.2: Taxonomic and functional composition of samples collected from 10 patients who received FMT from donor B and a patient on placebo treatment using shotgun metagenomics. The UPGMA tree of Bray-Curtis distances based on taxonomic composition (**A**) and microbial gene families (**B**). The colours of the inner line show samples collected from patients prior and post FMT (light and dark green respectively) or placebo (light and dark blue respectively) and donor B (purple). The outer layer shows the taxonomic composition of samples at the phylum-level (**A**) and the percentage of annotated gene families (**B**). **C**. The PCoA of Aitchison distances based on the taxonomic composition of assigned reads. Dotted lines connect samples collected at week 0 and week 6 for each individual. **D**. The PcoA of Aitchison distances based on the composition of known gene families in each sample. The number of donor B's species (**E**) and microbial gene families (**F**) engrafted in $\geq$ three patients post-FMT.

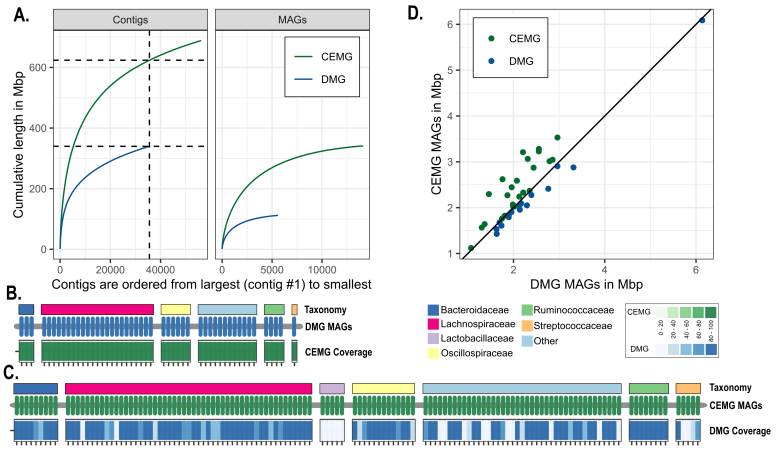

FIGURE A2.3: Comparison of culture-enriched (CEMG) and direct metagenomics (DMG) for a single donor B sample. **A**. The cumulative lengths of assembled contigs for the contigs and MAGs. **B**. DMG and CEMG coverage (percentage of MAG covered at least 1X) of the MAGs (n=49) assembled via DMG. The top colour bar shows the taxonomy of MAGs at the family-level. **B.** DMG and CEMG coverage of the MAGs (n=49) assembled via DMG. **C.** CEMG and DMG coverage for the MAGs (n=132) assembled via CEMG. **D**. Comparison of the genome size among homologous assembled MAGs in CEMG and DMG.

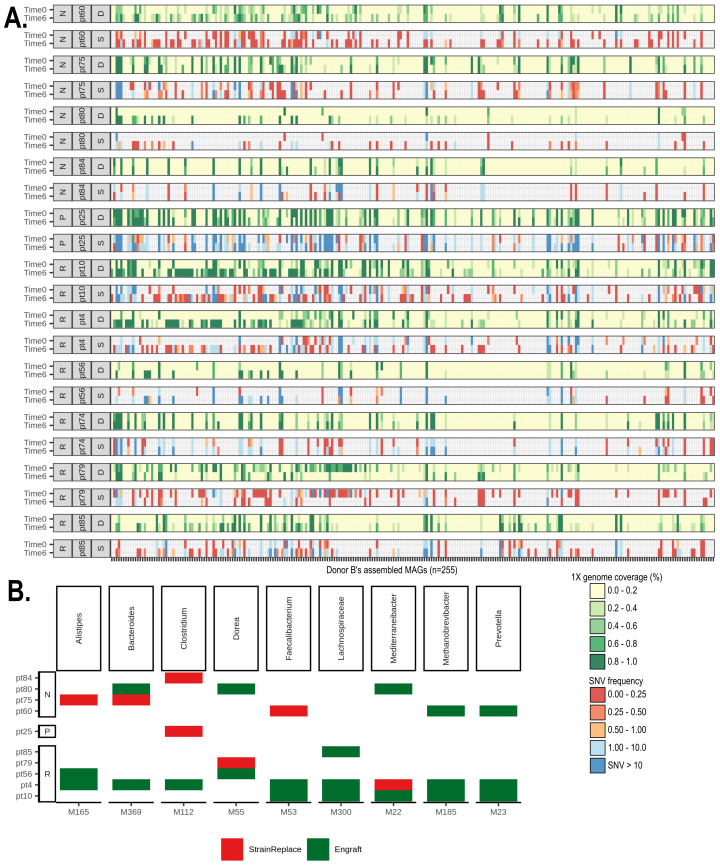

FIGURE A2.4: Tracking donor B MAGs after FMT. **A.** The genomic coverage (percentage of 1X; D) and SNV frequencies (S) of donor B MAGs (n=255) in samples collected from patients prior and post FMT or placebo treatment. Non-responder, placebo, and responder patients are labelled as N, P, R respectively. **B.** Commonly engrafted MAGs in $\geq$ three patient post FMT.

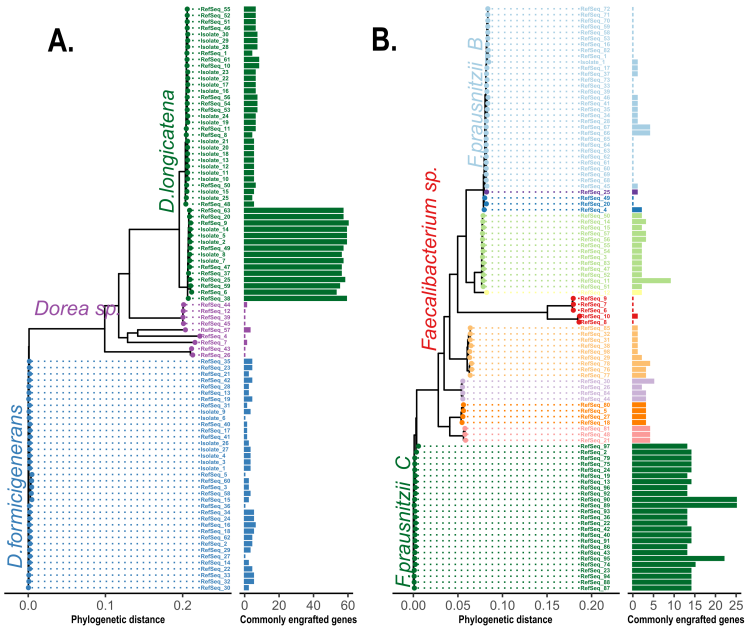

FIGURE A2.5: The commonly engrafted genes are strain-specific. A phylogeny of available strains in NCBI (**RefSeq #**) as well as Surrette lab whole genome collection (**Isolate #**) constructed for *Dorea sp.* (**A**) and *Faecalibacterium sp.* (**B**). The number of commonly engrafted genes identified in each genome are shown in **A.** *Dorea sp.* (n=95), **B.** *Faecalibacterium sp.* (n=99) phylogenies.

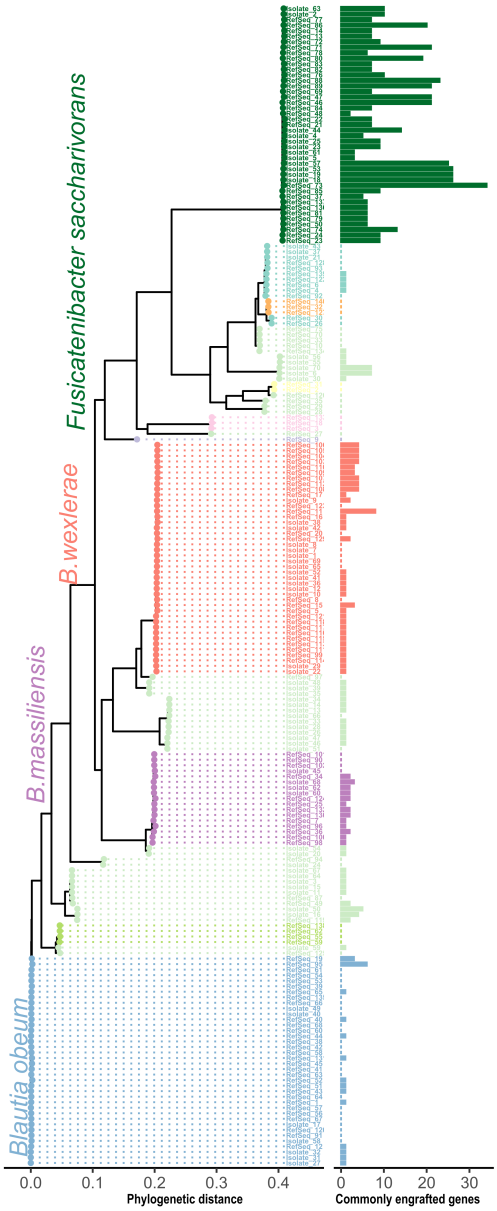

FIGURE A2.6: The commonly engrafted genes are strain-specific. A phylogeny of available strains in NCBI (**RefSeq #**) as well as Surrette lab whole genome collection (**Isolate #**) constructed for *Blautia sp.*. The number of commonly engrafted genes identified in each genome are shown in *Blautia sp.* (n=210) phylogeny. A clade of *Fusicatenubacter saccharivorans* genomes that contains most of the commonly engrafted genes identified in this collection using GTDB-Tk Chaumeil et al. 2019.

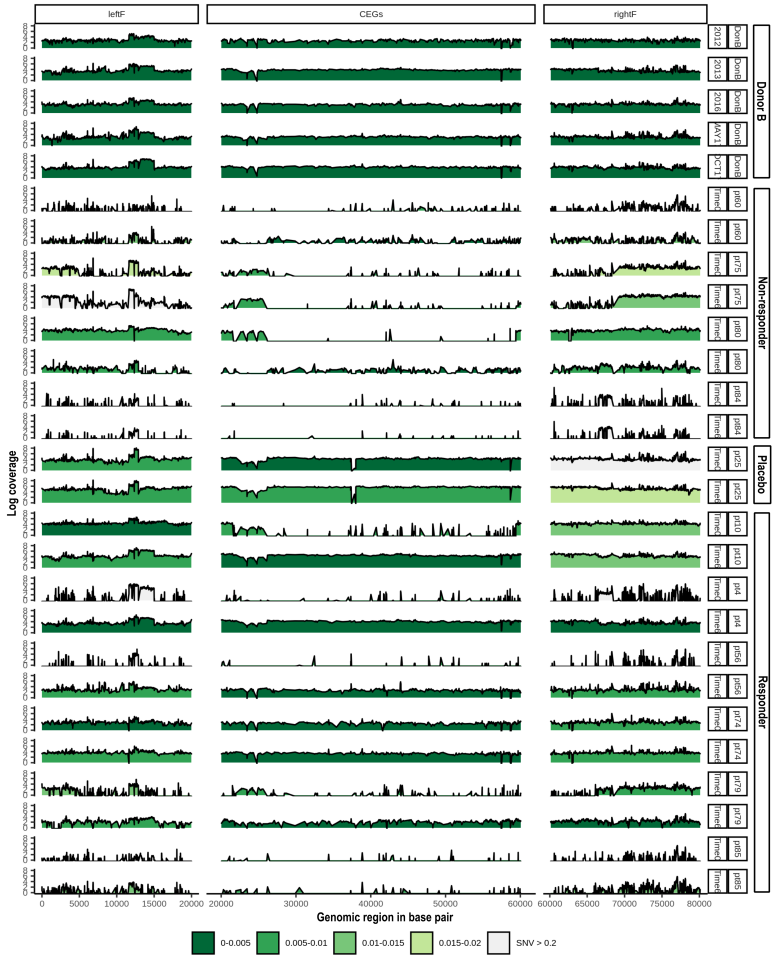

FIGURE A2.7: The genomic coverage and variability of the commonly engrafted gene (CEGs) cluster as well as flanking regions in *Fusicatenubacter saccharivorans*. Gene clusters are coloured based on their variability relative to stable microbial base positions stable in donor B samples. The frequency of SNVs calculated in each gene cluster relative to donor B samples and shown from less variable (green) to more variable (white) in donor B and patient samples. For example, the genomic coverage and SNV frequency of this 80 kbp region from patient 4, who responded to FMT, became similar to donor B following FMT.

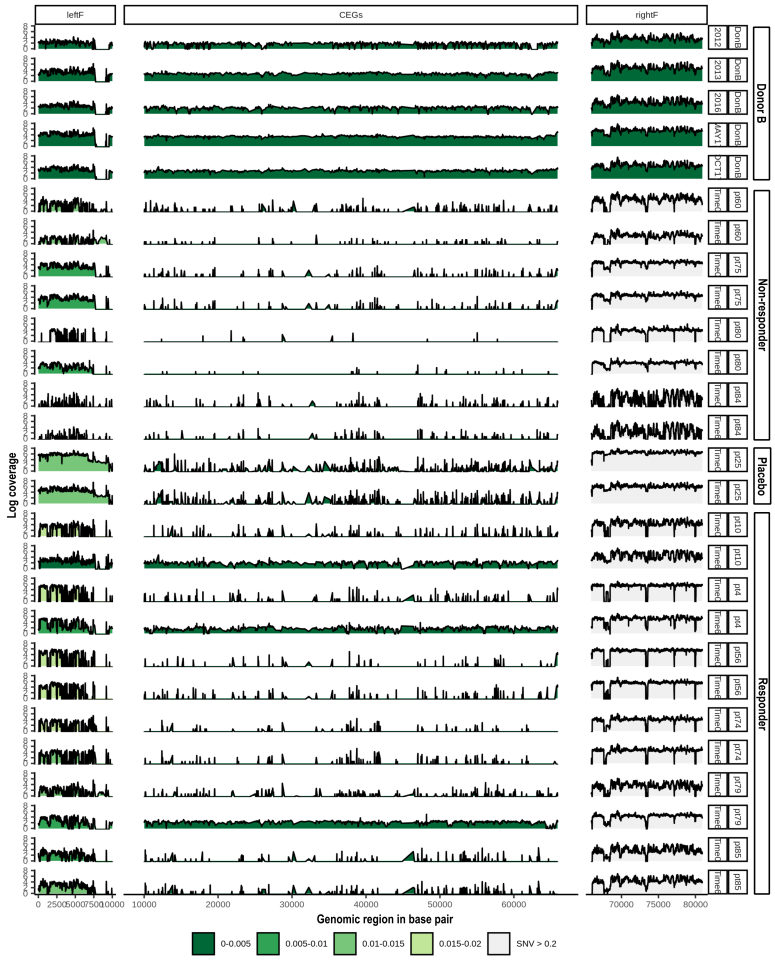

FIGURE A2.8: The genomic coverage and variability of the commonly engrafted gene (CEGs) cluster as well as flanking regions in *Faecalibacterium prausnitzii*. Gene clusters are coloured based on their variability relative to stable microbial base positions stable in donor B samples. The frequency of SNVs calculated in each gene cluster relative to donor B samples and shown from less variable (green) to more variable (white) in donor B and patient samples. For example, the identified commonly engrafted genes in patient 4 accompany a variable right flanking region following FMT.

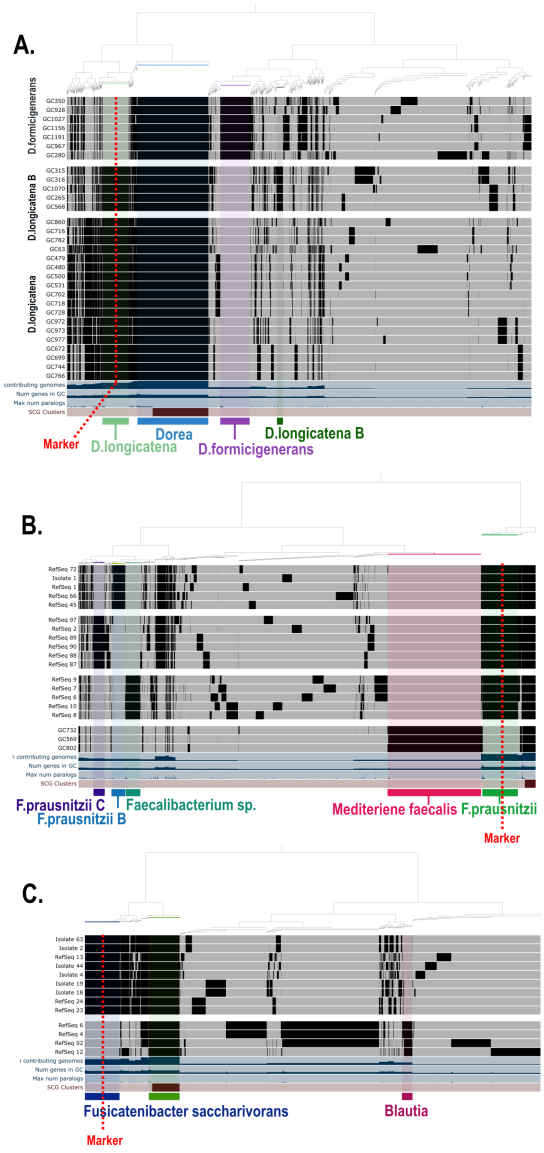

FIGURE A2.9: Building species-specific markers for **A.** *D. longicatena*, **B.** *F. prausnitzii*, and **C.** *F. saccharivorans*. Gene clusters are aligned across the pangenome, and 50 kb core-specific regions are selected as markers for each species. Each row shows a single genome and the aligned gene clusters in black. The gene clusters are shown in the x-axis for each collection, and core-specific regions are labelled in different colours. The red dotted line for each species shows the 50 kb marker.

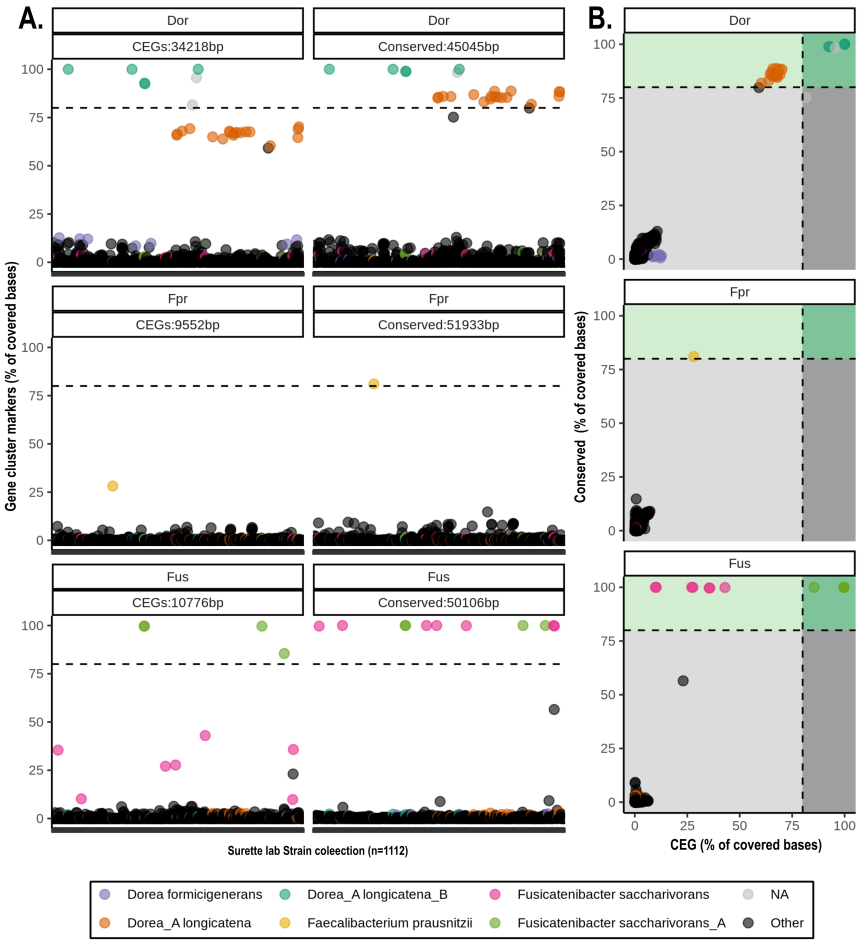

FIGURE A2.10: Validating the accuracy of strain- and species-specific markers using a diverse collection of 1112 human gut bacterial whole-genome sequences (WGS). Shotgun reads from each WGS were mapped to both markers in each representative strain of commonly engrafted genes. **A.** Comparison of strain-specific (CEGs) and species-specific (Conserved) markers in two panels. Each dot shows a single WGS, and the y-axis shows the percentage of 1X coverage. **B.** Comparison of strain-specific (CEGs) and species-specific (Conserved) markers in a single plot.

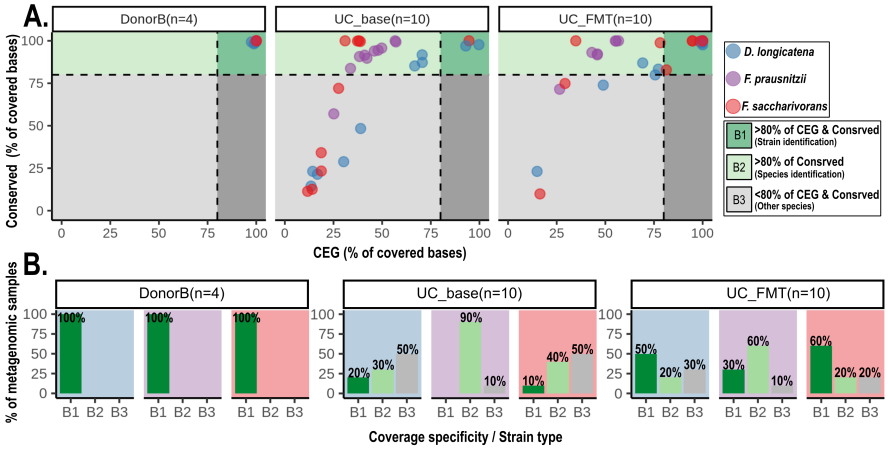

FIGURE A2.11: Tracking the representative strains of commonly engrafted genes in metagenomic samples using strain and species-specific markers. The specificity of *D. longicatena*, *F. prausnitzii*, and *F. saccharivorans* representative strains compared across metagenomic samples from this study (top row in each figure; 4 samples from donor B(n=1), 10 samples before FMT(n=10), and 10 samples following FMT (n=10)). **A.** Comparison of a conserved (species-specific) vs. commonly engrafted gene (strain-specific) cluster for each strain within each metagenomic sample. Each dot represents one genome in a metagenomic sample. **B.** The classified genomes from a metagenomic sample based on conserved and commonly engrafted gene's coverage percentage. Genomes with CEG=commonly engrafted genes and conserved gene cluster coverage $\geq$ 80% (dark green) and those with conserved coverage $\geq$ 80% (light green) in a metagenomic sample are labelled as B1 (strain-specific) and B2 (species-specific) respectively. The genomes with conserved region coverage $<$ 80% are labelled as B3 (other species). rsCEGs= representative strain of commonly engrafted genes from Figure 3.6. As we expected, all the samples collected from donor B (n=4) contain the representative strains ($\geq$80% of both conserved and commonly engrafted gene markers. Samples from UC patients post-FMT showed increased percentage of rsCEG strains compared to samples collected prior to FMT. *D. longicatena*, *F. prausnitzii*, and *F. saccharivorans* species that are present in 30%, 90%, and 40% of patients respectively are replaced by donor B strains post-FMT. While the percentage of species-level (other strain) detection reduced post-FMT, the percentage of *D. longicatena*, *F. prausnitzii*, and *F. saccharivorans* representative strains increased to 50%, 30%, and 60% respectively.

Bibliography

Abbas-Egbariya, H., Haberman, Y., Braun, T., Hadar, R., Denson, L., Gal-Mor, O., and Amir, A. (2022). Meta-analysis defines predominant shared microbial responses in various diseases and a specific inflammatory bowel disease signal. *Genome Biology* 23(1), 1–23.

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., and Neuhaus, K. (2021). Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *mSphere* 6(1), e01202–20.

Aggarwala, V., Mogno, I., Li, Z., Yang, C., Britton, G. J., Chen-Liaw, A., Mitcham, J., Bongers, G., Gevers, D., Clemente, J. C., et al. (2021). Precise quantification of bacterial strains after fecal microbiota transplantation delineates long-term engraftment and explains outcomes. *Nature Microbiology*, 1–10.

Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., et al. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research* 48(D1), D517–D525.

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., and Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature* 568(7753), 499–504.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., et al. (2021). A unified catalog

of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* 39(1), 105–114.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods* 11(11), 1144.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17), 3389–3402.

Ames, S. K., Hysom, D. A., Gardner, S. N., Lloyd, G. S., Gokhale, M. B., and Allen, J. E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* 29(18), 2253–2260.

Angelberger, S., Reinisch, W., Makristathis, A., Lichtenberger, C., Dejaco, C., Papay, P., Novacek, G., Trauner, M., Loy, A., and Berry, D. (2013). Temporal bacterial community dynamics vary among ulcerative colitis patients after fecal microbiota transplantation. *The American Journal of Gastroenterology* 108(10), 1620.

Arboleya, S., Ang, L., Margolles, A., Yiyuan, L., Dongya, Z., Liang, X., Solis, G., Fernandez, N., Clara, G., and Gueimonde, M. (2012). Deep 16S rRNA metagenomics and quantitative PCR analyses of the premature infant fecal microbiota. *Anaerobe* 18(3), 378–380.

Avrani, S., Schwartz, D. A., and Lindell, D. (2012). Virus-host swinging party in the oceans: Incorporating biological complexity into paradigms of antagonistic coexistence. *Mobile Genetic Elements* 2(2), 88–95.

Axelrad, J. E., Joelson, A., Green, P. H., Lawlor, G., Lichtiger, S., Cadwell, K., and Lebwohl, B. (2018). Enteric infections are common in patients with flares of inflammatory bowel disease. *The American Journal of Gastroenterology* 113(10), 1530.

Bagdasarian, N., Rao, K., and Malani, P. N. (2015). Diagnosis and treatment of *Clostridium difficile* in adults: a systematic review. *Jama* 313(4), 398–408.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19(5), 455–477.

Bartram, A. K., Lynch, M. D., Stearns, J. C., Moreno-Hagelsieb, G., and Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology* 77(11), 3846–3852.

Bastiaanssen, T. F., Cowan, C. S., Claesson, M. J., Dinan, T. G., and Cryan, J. F. (2019). Making sense of the microbiome in psychiatry. *International Journal of Neuropsychopharmacology* 22(1), 37–52.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., et al. (2004). The Pfam protein families database. *Nucleic Acids Research* 32(suppl_1), D138–D141.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bazinet, A. L. and Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics* 13(1), 1–13.

Beghini, F., McIver, L. J., Blanco-Miguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 10, e65088.

Benchimol, E. I., Guttmann, A., Griffiths, A. M., Rabeneck, L., Mack, D. R., Brill, H., Howard, J., Guan, J., and To, T. (2009). Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* 58(11), 1490–1497.

Bennet, J. and Brinkman, M. (1989). Treatment of ulcerative colitis by implantation of normal colonic flora. *The Lancet* 333(8630), 164.

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., Zhang, M., Oh, P. L., Nehrenberg, D., Hua, K., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences (USA)* 107(44), 18933–18938.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8(1), 1–22.

Berger, B., Porta, N., Foata, F., Grathwohl, D., Delley, M., Moine, D., Charpagne, A., Siegwald, L., Descombes, P., Alliet, P., et al. (2020). Linking human milk oligosaccharides, infant fecal community types, and later risk to require antibiotics. *Mbio* 11(2), e03196–19.

Bernstein, C. N., Burchill, C., Targownik, L. E., Singh, H., and Roos, L. L. (2019). Events within the first year of life, but not the neonatal period, affect risk for later development of inflammatory bowel diseases. *Gastroenterology* 156(8), 2190–2197.

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., Lieber, A. D., Wu, F., Perez-Perez, G. I., Chen, Y., et al. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* 8(343), 343ra82–343ra82.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), 2114–2120.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37(8), 852–857.

Borody, T. J., George, L., Andrews, P., Brandl, S., Noonan, S., Cole, P., Hyland, L., Morgan, A., Maysey, J., and Moore-Jones, D. (1989). Bowel-flora alteration: a potential cure for inflammatory bowel disease and irritable bowel syndrome? *Medical Journal of Australia* 150(10), 604–604.

Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., Cesare, M. de, Oikkonen, L. E., Parkes, D., Freeman, C., Dhalla, F., Patel, S. Y., et al. (2019). Sequencing of human genomes with nanopore technology. *Nature Communications* 10(1), 1–9.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35(8), 725–731.

Brady, A. and Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* 6(9), 673–676.

Brenner, D. J., Fanning, G., Skerman, F., and Falkow, S. (1972). Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *Journal of Bacteriology* 109(3), 953–965.

Britton, R. A. and Young, V. B. (2014). Role of the intestinal microbiota in resistance to colonization by *Clostridium difficile*. *Gastroenterology* 146(6), 1547–1553.

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., et al. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology* 15(1), 1–14.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., and Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523(7559), 208–211.

Brown, E. M., Arellano-Santoyo, H., Temple, E. R., Costliow, Z. A., Pichaud, M., Hall, A. B., Liu, K., Durney, M. A., Gu, X., Plichta, D. R., et al. (2021). Gut microbiome ADP-ribosyltransferases are widespread phage-encoded fitness factors. *Cell Host & Microbe* 29(9), 1351–1365.

Byndloss, M. X., Olsan, E. E., Rivera-Chávez, F., Tiffany, C. R., Cevallos, S. A., Lokken, K. L., Torres, T. P., Byndloss, A. J., Faber, F., Gao, Y., et al. (2017). Microbiota-activated PPAR-γ signaling inhibits dysbiotic Enterobacteriaceae expansion. *Science* 357(6351), 570–575.

Calkins, B. M. (1989). A meta-analysis of the role of smoking in inflammatory bowel disease. *Digestive Diseases and Sciences* 34(12), 1841–1854.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13(7), 581–583.

Campbell, E. L., Bruyninckx, W. J., Kelly, C. J., Glover, L. E., McNamee, E. N., Bowers, B. E., Bayless, A. J., Scully, M., Saeedi, B. J., Golden-Mason, L., et al. (2014). Transmigrating neutrophils shape the mucosal microenvironment through localized oxygen depletion to influence resolution of inflammation. *Immunity* 40(1), 66–77.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* 38(12), 5825–5829.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., et al. (2010a). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5), 335–336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2010b). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences (USA)*, 201000080.

Chaisson, M. J., Wilson, R. K., and Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics* 16(11), 627–640.

Chatterjee, A., Modarai, M., Naylor, N. R., Boyd, S. E., Atun, R., Barlow, J., Holmes, A. H., Johnson, A., and Robotham, J. V. (2018). Quantifying drivers of antibiotic resistance in humans: a systematic review. *The Lancet Infectious Diseases* 18(12), e368–e378.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36(6), 1925–1927.

Chen, J., Quiles-Puchalt, N., Chiang, Y. N., Bacigalupe, R., Fillol-Salom, A., Chee, M. S. J., Fitzgerald, J. R., and Penadés, J. R. (2018). Genome hypermobility by lateral transduction. *Science* 362(6411), 207–212.

Cheng, F., Huang, Z., Wei, W., and Li, Z. (2021). Fecal microbiota transplantation for Crohn's disease: A systematic review and meta-analysis. *Techniques in Coloproctology* 25(5), 495–504.

Chu, N. D., Crothers, J. W., Nguyen, L. T., Kearney, S. M., Smith, M. B., Kassam, Z., Collins, C., Xavier, R., Moses, P. L., and Alm, E. J. (2021). Dynamic colonization of microbes and their functions after fecal microbiota transplantation for inflammatory bowel disease. *Mbio* 12(4), e00975–21.

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcón, Ó., et al. (2015). The microbiome of uncontacted Amerindians. *Science Advances* 1(3), e1500183.

Coelho, L. P., Alves, R., Del Rio, A. R., Myers, P. N., Cantalapiedra, C. P., Giner-Lamia, J., Schmidt, T. S., Mende, D. R., Orakov, A., Letunic, I., et al. (2022). Towards the biogeography of prokaryotic genes. *Nature* 601(7892), 252–256.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., and Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42(D1), D633–D642.

Collins, S. M., Surette, M., and Bercik, P. (2012). The interplay between the intestinal microbiota and the brain. *Nature Reviews Microbiology* 10(11), 735.

Conway, T. C. and Bromage, A. J. (2011). Succinct data structures for assembling large genomes. *Bioinformatics* 27(4), 479–486.

Cornuault, J. K., Petit, M.-A., Mariadassou, M., Benevides, L., Moncaut, E., Langella, P., Sokol, H., and De Paepe, M. (2018). Phages infecting *Faecalibacterium prausnitzii* belong to novel viral genera that help to decipher intestinal viromes. *Microbiome* 6(1), 1–14.

Costello, S. P., Hughes, P. A., Waters, O., Bryant, R. V., Vincent, A. D., Blatchford, P., Katsikeros, R., Makanyanga, J., Campaniello, M. A., Mavrangelos, C., et al. (2019). Effect of fecal microbiota transplantation on 8-week remission in patients with ulcerative colitis: a randomized clinical trial. *Jama* 321(2), 156–164.

Coyne, M. J., Zitomersky, N. L., McGuire, A. M., Earl, A. M., and Comstock, L. E. (2014). Evidence of extensive DNA transfer between bacteroidales species within the human gut. *MBio* 5(3), e01305–14.

Crohn′s and Colitis Canada (2018). The Impact of Inflammatory Bowel Disease in Canada. *the scientific community to Crohn′s and Colitis Canada*. URL: `https://crohnsandcolitis.ca/Crohns_and_Colitis/documents/reports/2018-Impact-Report-LR.pdf`.

Crothers, J., Kassam, Z., Smith, M., Phillips, M., Vo, E., Velez, M., Cohn, A. H., Fortner, K., Guerra, R. D. R., Chu, N., et al. (2018). A double-blind, randomized, placebo-control pilot trial of fecal microbiota transplantation capsules from rationally selected donors in active ulcerative colitis. *Gastroenterology* 154(6), S–1050.

Cui, B., Feng, Q., Wang, H., Wang, M., Peng, Z., Li, P., Huang, G., Liu, Z., Wu, P., Fan, Z., et al. (2015). Fecal microbiota transplantation through mid-gut for refractory C rohn's disease: Safety, feasibility, and efficacy trial results. *Journal of Gastroenterology and Hepatology* 30(1), 51–58.

Danese, S. (2012). New therapies for inflammatory bowel disease: from the bench to the bedside. *Gut* 61(6), 918–932.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505(7484), 559.

De Filippis, F., Pasolli, E., and Ercolini, D. (2020). Newly explored Faecalibacterium diversity is connected to age, lifestyle, geography, and disease. *Current Biology* 30(24), 4932–4943.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences (USA)* 107(33), 14691–14696.

De Groot, P. F., Frissen, M., De Clercq, N., and Nieuwdorp, M. (2017). Fecal microbiota transplantation in metabolic syndrome: history, present and future. *Gut Microbes* 8(3), 253–267.

De Palma, G., Lynch, M. D., Lu, J., Dang, V. T., Deng, Y., Jury, J., Umeh, G., Miranda, P. M., Pigrau Pastor, M., Sidani, S., et al. (2017). Transplantation of fecal microbiota from patients with irritable bowel syndrome alters gut function and behavior in recipient mice. *Science Translational Medicine* 9(379), eaaf6397.

De Souza, H. S. and Fiocchi, C. (2016). Immunopathogenesis of IBD: current state of the art. *Nature Reviews Gastroenterology & Hepatology* 13(1), 13–27.

Deatherage, D. E. and Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. In: *Engineering and analyzing multicellular systems*. Springer, 165–188.

DeFilipp, Z., Bloom, P. P., Torres Soto, M., Mansour, M. K., Sater, M. R., Huntley, M. H., Turbett, S., Chung, R. T., Chen, Y.-B., and Hohmann, E. L. (2019). Drug-resistant E. coli bacteremia transmitted by fecal microbiota transplant. *New England Journal of Medicine* 381(21), 2043–2050.

Derakhshani, H., Bernier, S. P., Marko, V. A., and Surette, M. G. (2020). Completion of draft bacterial genomes by long-read sequencing of synthetic genomic pools. *BMC Genomics* 21(1), 1–11.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* 72(7), 5069–5072.

Deschasaux, M., Bouter, K. E., Prodan, A., Levin, E., Groen, A. K., Herrema, H., Tremaroli, V., Bakker, G. J., Attaye, I., Pinto-Sietsma, S.-J., et al. (2018). Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nature Medicine* 24(10), 1526–1531.

Dethlefsen, L. and Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences (USA)* 108(Supplement 1), 4554–4561.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., and Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology* 10(8), 1–16.

Dijk, L. R. van, Walker, B. J., Straub, T. J., Worby, C. J., Grote, A., Schreiber, H. L., Anyansi, C., Pickering, A. J., Hultgren, S. J., Manson, A. L., et al. (2022). StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities. *Genome Biology* 23(1), 1–27.

Dijkshoorn, L., Ursing, B., and Ursing, J. (2000). Strain, clone and species: comments on three basic concepts of bacteriology. *Journal of Medical Microbiology* 49(5), 397–401.

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences (USA)* 107(26), 11971–11975.

Dominguez-Bello, M. G., Blaser, M. J., Ley, R. E., and Knight, R. (2011). Development of the human gastrointestinal microbiota and insights from high-throughput sequencing. *Gastroenterology* 140(6), 1713–1719.

Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R., and Blaser, M. J. (2019). Role of the microbiome in human development. *Gut* 68(6), 1108–1114.

Draper, L. A., Ryan, F. J., Smith, M. K., Jalanka, J., Mattila, E., Arkkila, P., Ross, R. P., Satokari, R., and Hill, C. (2018). Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation. *Microbiome* 6(1), 1–9.

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* 5(1), 1–11.

Dutilh, B. E., Huynen, M. A., and Strous, M. (2009). Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics* 25(21), 2878–2881.

Eaden, J., Abrams, K., and Mayberry, J. (2001). The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut* 48(4), 526–535.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34(14), 2371–2375.

Edwards, R. A., Vega, A. A., Norman, H. M., Ohaeri, M., Levi, K., Dinsdale, E. A., Cinek, O., Aziz, R. K., McNair, K., Barr, J. J., et al. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology* 4(10), 1727–1736.

Ehrlich, S. D., Consortium, M., et al. (2011). MetaHIT: The European Union Project on metagenomics of the human intestinal tract. In: *Metagenomics of the human body.* Springer, 307–316.

Ekbom, A., Helmick, C., Zack, M., and Adami, H.-O. (1990). Ulcerative colitis and colorectal cancer: a population-based study. *New England Journal of Medicine* 323(18), 1228–1233.

Enav, H. and Ley, R. E. (2021). SynTracker: a synteny based tool for tracking microbial strains. *BioRxiv*. URL: `https://doi.org/10.1101/2021.10.06.463341`.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., and Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3, e1319.

Fehily, S. R., Basnayake, C., Wright, E. K., and Kamm, M. A. (2021). Fecal microbiota transplantation therapy in Crohn's disease: Systematic review. *Journal of Gastroenterology and Hepatology* 36(10), 2672–2686.

Fernandes, F., Pereira, L., and Freitas, A. T. (2009). CSA: an efficient algorithm to improve circular DNA multiple alignment. *BMC Bioinformatics* 10(1), 1–13.

Finegold, S. M., Attebery, H. R., and Sutter, V. L. (1974). Effect of diet on human fecal flora: comparison of Japanese and American diets. *The American Journal of Clinical Nutrition* 27(12), 1456–1469.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223), 496–512.

Floch, M. H. (2015). Intestinal microbiota metabolism of a prebiotic to treat hepatic encephalopathy. *Clinical Gastroenterology and Hepatology* 13(1), 209.

Ford, A. C., Harris, L. A., Lacy, B. E., Quigley, E. M., and Moayyedi, P. (2018). Systematic review with meta-analysis: the efficacy of prebiotics, probiotics, synbiotics and antibiotics in irritable bowel syndrome. *Alimentary Pharmacology & Therapeutics* 48(10), 1044–1060.

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., Dunn, M., Mkandawire, T. T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology* 37(2), 186–192.

Fox, G. E., Pechman, K. R., and Woese, C. R. (1977). Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *International Journal of Systematic and Evolutionary Microbiology* 27(1), 44–57.

Frank, D. N., St. Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences (USA)* 104(34), 13780–13785.

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* 15(11), 962–968.

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., et al. (2019). Gut microbiome

structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* 4(2), 293–305.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., et al. (1995). The minimal gene complement of Mycoplasma genitalium. *Science* 270(5235), 397–404.

Fritz, A., Hofmann, P., Majda, S., Dahms, E., Droge, J., Fiedler, J., Lesker, T. R., Belmann, P., DeMaere, M. Z., Darling, A. E., et al. (2019). CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 7(1), 1–12.

Fuentes, S., Rossen, N. G., Spek, M. J. van der, Hartman, J. H., Huuskonen, L., Korpela, K., Salojärvi, J., Aalvink, S., Vos, W. M. de, D'Haens, G. R., et al. (2017). Microbial shifts and signatures of long-term remission in ulcerative colitis after faecal microbiota transplantation. *The ISME journal* 11(8), 1877–1889.

Fuentes, S., Van Nood, E., Tims, S., Heikamp-de Jong, I., Ter Braak, C. J., Keller, J. J., Zoetendal, E. G., and De Vos, W. M. (2014). Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent *Clostridium difficile* infection. *The ISME journal* 8(8), 1621.

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* 504(7480), 446.

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences (USA)* 109(2), 594–599.

GS, B., AJ, K., et al. (1958). Fecal enema as an adjunct in the treatment of pseudomembranous enterocolitis. *Surgery* 44(5), 854–859.

Guerin, E., Shkoporov, A., Stockdale, S. R., Clooney, A. G., Ryan, F. J., Sutton, T. D., Draper, L. A., Gonzalez-Tortuero, E., Ross, R. P., and Hill, C. (2018). Biology and

taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host & Microbe* 24(5), 653–664.

Haifer, C., Luu, L. D. W., Paramsothy, S., Borody, T. J., Leong, R. W., and Kaakoush, N. O. (2022a). Microbial determinants of effective donors in faecal microbiota transplantation for UC. *Gut*.

Haifer, C., Paramsothy, S., Kaakoush, N. O., Saikal, A., Ghaly, S., Yang, T., Luu, L. D. W., Borody, T. J., and Leong, R. W. (2022b). Lyophilised oral faecal microbiota transplantation for ulcerative colitis (LOTUS): a randomised, double-blind, placebo-controlled trial. *The Lancet Gastroenterology & Hepatology* 7(2), 141–151.

Hamilton, M. K., Ronveaux, C. C., Rust, B. M., Newman, J. W., Hawley, M., Barile, D., Mills, D. A., and Raybould, H. E. (2017). Prebiotic milk oligosaccharides prevent development of obese phenotype, impairment of gut permeability, and microbial dysbiosis in high fat-fed mice. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 312(5), G474–G487.

Hamilton, M. J., Weingarden, A. R., Unno, T., Khoruts, A., and Sadowsky, M. J. (2013). High-throughput DNA sequence analysis reveals stable engraftment of gut microbiota following transplantation of previously frozen fecal bacteria. *Gut Microbes* 4(2), 125–135.

Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F. M., Briggs, J., et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn's disease in ATG16L1. *Nature Genetics* 39(2), 207–211.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 5(10), R245–R249.

He, R., Li, P., Wang, J., Cui, B., Zhang, F., and Zhao, F. (2022). The interplay of gut microbiota between donors and recipients determines the efficacy of fecal microbiota transplantation. *Gut Microbes* 14(1), 2100197.

Hibbing, M. E., Fuqua, C., Parsek, M. R., and Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology* 8(1), 15–25.

Hilt, E. E., McKinley, K., Pearce, M. M., Rosenfeld, A. B., Zilliox, M. J., Mueller, E. R., Brubaker, L., Gai, X., Wolfe, A. J., and Schreckenberger, P. C. (2014). Urine is not sterile: use of enhanced urine culture techniques to detect resident bacterial flora in the adult female bladder. *Journal of Clinical Microbiology* 52(3), 871–876.

Hindryckx, P., Jairath, V., and D'haens, G. (2016). Acute severe ulcerative colitis: from pathophysiology to clinical management. *Nature Reviews Gastroenterology & Hepatology* 13(11), 654–664.

Ho, S., Pothoulakis, C., and Wai Koon, H. (2013). Antimicrobial peptides and colitis. *Current pharmaceutical design* 19(1), 40–47.

Hooper, L. V., Littman, D. R., and Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science* 336(6086), 1268–1273.

Huang, X., Wang, J., Aluru, S., Yang, S.-P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. *Genome Research* 13(9), 2164–2170.

Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H., and Soo, R. M. (2021). Prokaryotic taxonomy and nomenclature in the age of big sequence data. *the ISME Journal* 15(7), 1879–1892.

Hugon, P., Dufour, J.-C., Colson, P., Fournier, P.-E., Sallah, K., and Raoult, D. (2015). A comprehensive repertoire of prokaryotic species identified in human beings. *The Lancet Infectious Diseases* 15(10), 1211–1219.

Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.-P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C. A., Gassull, M., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411(6837), 599–603.

Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research* 21(9), 1552–1560.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11(1), 1–11.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5(3), 299–314.

Imielinski, M., Baldassano, R. N., Griffiths, A., Russell, R. K., Annese, V., Dubinsky, M., Kugathasan, S., Bradfield, J. P., Walters, T. D., Sleiman, P., et al. (2009). Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nature Genetics* 41(12), 1335–1340.

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

James, S. L., Christophersen, C. T., Bird, A. R., Conlon, M. A., Rosella, O., Gibson, P. R., and Muir, J. G. (2015). Abnormal fibre usage in UC in remission. *Gut* 64(4), 562–570.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* 10(1), 1–11.

Kamada, N., Chen, G. Y., Inohara, N., and Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. *Nature Immunology* 14(7), 685–690.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7, e7359.

Kappelman, M. D., Rifas–Shiman, S. L., Kleinman, K., Ollendorf, D., Bousvaros, A., Grand, R. J., and Finkelstein, J. A. (2007). The prevalence and geographic distribution of Crohn's disease and ulcerative colitis in the United States. *Clinical Gastroenterology and Hepatology* 5(12), 1424–1429.

Karcher, N., Nigro, E., Michal, P., Blanco-Miguez, A., Ciciani, M., Manghi, P., Zolfo, M., Cumbo, F., Manara, S., Golzato, D., et al. (2021). Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biology* 22(1), 1–24.

Karkman, A., Pärnänen, K., and Larsson, D. (2019). Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nature Communications* 10(1), 1–8.

Kassam, Z., Hundal, R., Marshall, J. K., and Lee, C. H. (2012). Fecal transplant via retention enema for refractory or recurrent *Clostridium difficile* infection. *Archives of Internal Medicine* 172(2), 191–193.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14), 3059–3066.

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

Khan, K. J., Ullman, T. A., Ford, A. C., Abreu, M. T., Abadir, A., Marshall, J. K., Talley, N. J., and Moayyedi, P. (2011). Antibiotic therapy in inflammatory bowel disease:

a systematic review and meta-analysis. *Official journal of the American College of Gastroenterology| ACG* 106(4), 661–673.

Khanna, S., Pardi, D. S., Aronson, S. L., Kammer, P. P., Orenstein, R., Sauver, J. L. S., Harmsen, W. S., and Zinsmeister, A. R. (2012). The epidemiology of community-acquired *Clostridium difficile* infection: a population-based study. *The American Journal of Gastroenterology* 107(1), 89.

Khanna, S., Shin, A., and Kelly, C. P. (2017a). Management of *Clostridium difficile* infection in inflammatory bowel disease: expert review from the clinical practice updates committee of the AGA institute. *Clinical Gastroenterology and Hepatology* 15(2), 166–174.

Khanna, S., Vazquez-Baeza, Y., González, A., Weiss, S., Schmidt, B., Muñiz-Pedrogo, D. A., Rainey, J. F., Kammer, P., Nelson, H., Sadowsky, M., et al. (2017b). Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease. *Microbiome* 5(1), 1–8.

Khoruts, A., Staley, C., and Sadowsky, M. J. (2021). Faecal microbiota transplantation for Clostridioides difficile: mechanisms and pharmacology. *Nature Reviews Gastroenterology & Hepatology* 18(1), 67–80.

Kirchgesner, J., Lemaitre, M., Carrat, F., Zureik, M., Carbonnel, F., and Dray-Spira, R. (2018). Risk of serious and opportunistic infections associated with treatment of inflammatory bowel diseases. *Gastroenterology* 155(2), 337–346.

Koonin, E. V., Makarova, K. S., Wolf, Y. I., and Krupovic, M. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nature Reviews Genetics* 21(2), 119–131.

Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., et al. (2012). Microbial culturomics: paradigm shift in the human gut microbiome study. *Clinical Microbiology and Infection* 18(12), 1185–1193.

Lagier, J.-C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S. I., Dubourg, G., et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology* 1(12), 1–8.

Lam, S., Bai, X., Shkoporov, A. N., Park, H., Wu, X., Lan, P., and Zuo, T. (2022). Roles of the gut virome and mycobiome in faecal microbiota transplantation. *The Lancet Gastroenterology & Hepatology*.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* 15(2), 141–161.

Lapaquette, P., Bringer, M.-A., and Darfeuille-Michaud, A. (2012). Defects in autophagy favour adherent-invasive *Escherichia coli* persistence within macrophages leading to increased pro-inflammatory response. *Cellular Microbiology* 14(6), 791–807.

Lau, J. T., Whelan, F. J., Herath, I., Lee, C. H., Collins, S. M., Bercik, P., and Surette, M. G. (2016). Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Medicine* 8(1), 72.

Lawley, T. D. and Walker, A. W. (2013). Intestinal colonization resistance. *Immunology* 138(1), 1–11.

Laxminarayan, R., Van Boeckel, T., Frost, I., Kariuki, S., Khan, E. A., Limmathurotsakul, D., Larsson, D. J., Levy-Hara, G., Mendelson, M., Outterson, K., et al. (2020). The Lancet Infectious Diseases Commission on antimicrobial resistance: 6 years later. *The Lancet Infectious Diseases* 20(4), e51–e60.

Lee, J. W. J., Plichta, D., Hogstrom, L., Borren, N. Z., Lau, H., Gregory, S. M., Tan, W., Khalili, H., Clish, C., Vlamakis, H., et al. (2021). Multi-omics reveal microbial determinants impacting responses to biologic therapies in inflammatory bowel disease. *Cell Host & Microbe* 29(8), 1294–1304.

Lee, K. A., Thomas, A. M., Bolte, L. A., Björk, J. R., Ruijter, L. K. de, Armanini, F., Asnicar, F., Blanco-Miguez, A., Board, R., Calbet-Llopart, N., et al. (2022). Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nature Medicine* 28(3), 535–544.

Lee, S. T., Kahn, S. A., Delmont, T. O., Shaiber, A., Esen, Ö. C., Hubert, N. A., Morrison, H. G., Antonopoulos, D. A., Rubin, D. T., and Eren, A. M. (2017). Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome* 5(1), 1–10.

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1), 127–128.

Levine, A., Boneh, R. S., and Wine, E. (2018). Evolving role of diet in the pathogenesis and treatment of inflammatory bowel diseases. *Gut* 67(9), 1726–1738.

Levine, J. M. and D'Antonio, C. M. (1999). Elton revisited: a review of evidence linking diversity and invasibility. *Oikos*, 15–26.

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., and Ettema, T. J. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology* 19(4), 225–240.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10), 1674–1676.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16), 2078–2079.

Li, S. S., Zhu, A., Benes, V., Costea, P. I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., Voigt, A. Y., et al. (2016). Durable coexistence of donor

and recipient strains after fecal microbiota transplantation. *Science* 352(6285), 586–589.

Liang, Y., Zhang, W., Tong, Y., and Chen, S. (2016). crAssphage is not associated with diarrhoea and has high genetic diversity. *Epidemiology & Infection* 144(16), 3549–3553.

Liao, X., Li, M., Zou, Y., Wu, F.-X., Wang, J., et al. (2019). Current challenges and solutions of de novo assembly. *Quantitative Biology* 7(2), 90–109.

Libertucci, J., Dutta, U., Kaur, S., Jury, J., Rossi, L., Fontes, M. E., Shajib, M. S., Khan, W. I., Surette, M. G., Verdu, E. F., et al. (2018). Inflammation-related differences in mucosa-associated microbiota and intestinal barrier function in colonic Crohn's disease. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 315(3), G420–G431.

Lin, H. and Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature Communications* 11(1), 1–11.

Lischer, H. E. and Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* 18(1), 1–12.

Liu, T.-C., Kern, J. T., Jain, U., Sonnek, N. M., Xiong, S., Simpson, K. F., VanDussen, K. L., Winkler, E. S., Haritunians, T., Malique, A., et al. (2021). Western diet induces Paneth cell defects through microbiome alterations and farnesoid X receptor and type I interferon activation. *Cell Host & Microbe* 29(6), 988–1001.

Llewellyn, S. R., Britton, G. J., Contijoch, E. J., Vennaro, O. H., Mortha, A., Colombel, J.-F., Grinspan, A., Clemente, J. C., Merad, M., and Faith, J. J. (2018). Interactions between diet and the intestinal microbiota alter intestinal permeability and colitis severity in mice. *Gastroenterology* 154(4), 1037–1046.

Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences (USA)* 113(21), 5970–5975.

Loeffler, C., Karlsberg, A., Martin, L. S., Eskin, E., Koslicki, D., and Mangul, S. (2020). Improving the usability and comprehensiveness of microbial databases. *BMC Biology* 18(1), 1–6.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489(7415), 220–230.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1(1), 2047–217X.

Lupp, C., Robertson, M. L., Wickham, M. E., Sekirov, I., Champion, O. L., Gaynor, E. C., and Finlay, B. B. (2007). Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. *Cell Host & Microbe* 2(2), 119–129.

Mahid, S. S., Minor, K. S., Soto, R. E., Hornung, C. A., and Galandiuk, S. (2006). Smoking and inflammatory bowel disease: a meta-analysis. In: *Mayo Clinic Proceedings*. Vol. 81. 11. Elsevier, 1462–1471.

Mahmoud, M., Zywicki, M., Twardowski, T., and Karlowski, W. M. (2019). Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. *Genomics* 111(1), 43–49.

Manrique, P., Bolduc, B., Walk, S. T., Oost, J. van der, Vos, W. M. de, and Young, M. J. (2016). Healthy human gut phageome. *Proceedings of the National Academy of Sciences (USA)* 113(37), 10400–10405.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17(1), 10–12.

Martinez-Medina, M. and Garcia-Gil, L. J. (2014). *Escherichia coli* in chronic inflammatory bowel diseases: An update on adherent invasive Escherichia coli pathogenicity. *World Journal of Gastrointestinal Pathophysiology* 5(3), 213.

Maslowski, K. M., Vieira, A. T., Ng, A., Kranich, J., Sierro, F., Yu, D., Schilter, H. C., Rolph, M. S., Mackay, F., Artis, D., et al. (2009). Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature* 461(7268), 1282–1286.

Maynard, C. L., Elson, C. O., Hatton, R. D., and Weaver, C. T. (2012). Reciprocal interactions of the intestinal microbiota and immune system. *Nature* 489(7415), 231–241.

McDonald, B., Zucoloto, A. Z., Yu, I.-L., Burkhard, R., Brown, K., Geuking, M. B., and McCoy, K. D. (2020). Programing of an intravascular immune firewall by the gut microbiota protects against pathogen dissemination during infection. *Cell Host & Microbe* 28(5), 660–668.

McMurdie, P. J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* 8(4), e61217.

Meyer, F., Lesker, T.-R., Koslicki, D., Fritz, A., Gurevich, A., Darling, A. E., Sczyrba, A., Bremges, A., and McHardy, A. C. (2021). Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nature Protocols* 16(4), 1785–1801.

Michail, S., Durbin, M., Turner, D., Griffiths, A. M., Mack, D. R., Hyams, J., Leleiko, N., Kenche, H., Stolfi, A., and Wine, E. (2012). Alterations in the gut microbiome of children with severe ulcerative colitis. *Inflammatory Bowel Diseases* 18(10), 1799–1808.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2015). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32(7), 1088–1090.

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences (USA)* 110(30), 12450–12455.

Mirsepasi-Lauridsen, H. C., Du, Z., Struve, C., Charbon, G., Karczewski, J., Krogfelt, K. A., Petersen, A. M., and Wells, J. M. (2016). Secretion of alpha-hemolysin by

Escherichia coli disrupts tight junctions in ulcerative colitis patients. *Clinical and Translational Gastroenterology* 7(3), e149.

Moayyedi, P. (2018). Update on fecal microbiota transplantation in patients with inflammatory bowel disease. *Gastroenterology & Hepatology* 14(5), 319.

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

Molodecky, N. A., Soon, S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., Benchimol, E. I., Panaccione, R., Ghosh, S., Barkema, H. W., et al. (2012). Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* 142(1), 46–54.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology* 13(9), 1–18.

Morrison, D. J. and Preston, T. (2016). Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* 7(3), 189–200.

Mottawea, W., Chiang, C.-K., Mühlbauer, M., Starr, A. E., Butcher, J., Abujamel, T., Deeke, S. A., Brandel, A., Zhou, H., Shokralla, S., et al. (2016). Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. *Nature Communications* 7(1), 1–14.

Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z., and Dominguez-Bello, M. G. (2015). The infant microbiome development: mom matters. *Trends in Molecular Medicine* 21(2), 109–117.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* 17(1), 1–9.

Mullen, K., Yasuda, K., Divers, T., and Weese, J. (2018). Equine faecal microbiota transplant: Current knowledge, proposed guidelines and future directions. *Equine Veterinary Education* 30(3), 151–160.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of Drosophila. *Science* 287(5461), 2196–2204.

Myles, I. A., Reckhow, J. D., Williams, K. W., Sastalla, I., Frank, K. M., and Datta, S. K. (2016). A method for culturing Gram-negative skin microbiota. *BMC Microbiology* 16(1), 1–6.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* 40(20), e155–e155.

Narula, N., Kassam, Z., Yuan, Y., Colombel, J.-F., Ponsioen, C., Reinisch, W., and Moayyedi, P. (2017). Systematic review and meta-analysis: fecal microbiota transplantation for treatment of active ulcerative colitis. *Inflammatory Bowel Diseases* 23(10), 1702–1709.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology* 39(4), 499–509.

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., and Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568(7753), 505–510.

New, F. N., Baer, B. R., Clark, A. G., Wells, M. T., and Brito, I. L. (2022). Collective effects of human genomic variation on microbiome function. *Scientific Reports* 12(1), 1–12.

Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., Panaccione, R., Ghosh, S., Wu, J. C., Chan, F. K., et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet* 390(10114), 2769–2778.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* 32(8), 822–828.

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., Kambal, A., Monaco, C. L., Zhao, G., Fleshner, P., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160(3), 447–460.

Nousias, O. and Montesanto, F. (2021). Metagenomic profiling of host-associated bacteria from 8 datasets of the red alga Porphyra purpurea with MetaPhlAn3. *Marine Genomics*, 100866.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research* 27(5), 824–834.

O'Toole, P. W. and Jeffery, I. B. (2015). Gut microbiota and aging. *Science* 350(6265), 1214–1215.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., et al. (2013). Package 'vegan'. *Community ecology package, version* 2(9), 1–295. URL: `https://cran.r-project.org/web/packages/vegan/index.html`.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2), 289–290.

Paramsothy, S., Kamm, M. A., Kaakoush, N. O., Walsh, A. J., Bogaerde, J. van den, Samuel, D., Leong, R. W., Connor, S., Ng, W., Paramsothy, R., et al. (2017). Multi-donor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *The Lancet* 389(10075), 1218–1228.

Paramsothy, S., Nielsen, S., Kamm, M. A., Deshpande, N. P., Faith, J. J., Clemente, J. C., Paramsothy, R., Walsh, A. J., Bogaerde, J. van den, Samuel, D., et al. (2019). Specific bacteria and metabolites associated with response to fecal microbiota transplantation in patients with ulcerative colitis. *Gastroenterology* 156(5), 1440–1454.

Paredes-Sabja, D., Shen, A., and Sorg, J. A. (2014). *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. *Trends in microbiology* 22(7), 406–416.

Park, S.-Y., Rao, C., Coyte, K. Z., Kuziel, G. A., Zhang, Y., Huang, W., Franzosa, E. A., Weng, J.-K., Huttenhower, C., and Rakoff-Nahoum, S. (2022). Strain-level fitness in the gut microbiome is an emergent property of glycans and a single metabolite. *Cell* 185(3), 513–529.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* 38(9), 1079–1086.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25(7), 1043–1055.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., and Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 2(11), 1533.

Peng, L., He, Z., Chen, W., Holzman, I. R., and Lin, J. (2007). Effects of butyrate on intestinal barrier function in a Caco-2 cell monolayer model of intestinal barrier. *Pediatric Research* 61(1), 37–41.

Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27(13), i94–i101.

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

Peters, L. A., Perrigoue, J., Mortha, A., Iuga, A., Song, W.-m., Neiman, E. M., Llewellyn, S. R., Di Narzo, A., Kidd, B. A., Telesco, S. E., et al. (2017). A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics* 49(10), 1437–1449.

Petersen, A. M., Nielsen, E. M., Litrup, E., Brynskov, J., Mirsepasi, H., and Krogfelt, K. A. (2009). A phylogenetic group of *Escherichia coli* associated with active left-sided inflammatory bowel disease. *BMC Microbiology* 9(1), 1–7.

Peterson, L. W. and Artis, D. (2014). Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nature Reviews Immunology* 14(3), 141–153.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., Hoeck, A. van, Wood, H. M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P. B., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature* 580(7802), 269–273.

Podlesny, D., Arze, C., Dörner, E., Verma, S., Dutta, S., Walter, J., and Fricke, W. F. (2022a). Metagenomic strain detection with SameStr: identification of a persisting core gut microbiota transferable by fecal transplantation. *Microbiome* 10(1), 1–15.

Podlesny, D., Durdevic, M., Paramsothy, S., Kaakoush, N. O., Hogenauer, C., Gorkiewicz, G., Walter, J., and Fricke, W. F. (2022b). Identification of clinical and ecological

determinants of strain engraftment after fecal microbiota transplantation using metagenomics. *Cell Reports Medicine*, 100711.

Poyet, M., Groussin, M., Gibbons, S. M., Avila-Pacheco, J., Jiang, X., Kearney, S. M., Perrotta, A., Berdy, B., Zhao, S., Lieberman, T., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature Medicine* 25(9), 1442–1452.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* 5(3), e9490.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490(7418), 55–60.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41(D1), D590–D596.

Quince, C., Nurk, S., Raguideau, S., James, R., Soyer, O. S., Summers, J. K., Limasset, A., Eren, A. M., Chikhi, R., and Darling, A. E. (2021). STRONG: metagenomics strain resolution on assembly graphs. *Genome Biology* 22(1), 1–34.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35(9), 833–844.

Quraishi, M. N., Widlak, M., Bhala, N. a., Moore, D., Price, M., Sharma, N., and Iqbal, T. (2017). Systematic review with meta-analysis: the efficacy of faecal microbiota transplantation for the treatment of recurrent and refractory *Clostridium difficile* infection. *Alimentary Pharmacology & Therapeutics* 46(5), 479–493.

Ragupathi, N. D., Sethuvel, D. M., Inbanathan, F., and Veeraraghavan, B. (2018). Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies. *New Microbes and New Infections* 21, 58–62.

Rajilić-Stojanović, M. and De Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS microbiology reviews* 38(5), 996–1047.

Rangan, P., Choi, I., Wei, M., Navarrete, G., Guen, E., Brandhorst, S., Enyati, N., Pasia, G., Maesincee, D., Ocon, V., et al. (2019). Fasting-mimicking diet modulates microbiota and promotes intestinal regeneration to reduce inflammatory bowel disease pathology. *Cell Reports* 26(10), 2704–2719.

Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual review of microbiology* 57(1), 369–394.

Rettedal, E. A., Gumpert, H., and Sommer, M. O. (2014). Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nature Communications* 5(1), 1–9.

Rhoads, A. and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics* 13(5), 278–289.

Rigottier Gois, L. (2013). Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *The ISME journal* 7(7), 1256–1261.

Rivera-Chávez, F., Zhang, L. F., Faber, F., Lopez, C. A., Byndloss, M. X., Olsan, E. E., Xu, G., Velazquez, E. M., Lebrilla, C. B., Winter, S. E., et al. (2016). Depletion of butyrate-producing *Clostridia* from the gut microbiota drives an aerobic luminal expansion of *Salmonella*. *Cell Host & Microbe* 19(4), 443–454.

Rodriguez-Beltran, J., Sorum, V., Toll-Riera, M., Vega, C. de la, Pe na-Miller, R., and San Millan, A. (2020). Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria. *Proceedings of the National Academy of Sciences (USA)* 117(27), 15755–15762.

Rossen, N. G., Fuentes, S., Spek, M. J. van der, Tijssen, J. G., Hartman, J. H., Duflou, A., Löwenberg, M., Brink, G. R. van den, Mathus-Vliegen, E. M., Vos, W. M. de, et al. (2015). Findings from a randomized controlled trial of fecal transplantation for patients with ulcerative colitis. *Gastroenterology* 149(1), 110–118.

Round, J. L. and Mazmanian, S. K. (2010). Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proceedings of the National Academy of Sciences (USA)* 107(27), 12204–12209.

Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernandez-Rodriguez, J., Clermont, O., Denamur, E., Rocha, E. P., and Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nature Microbiology* 6(3), 301–312.

Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* 4, e08490.

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V., et al. (2016). Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. *Cell* 167(6), 1469–1480.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75(23), 7537–7541.

Schmidt, T. S., Li, S. S., Maistrenko, O. M., Akkani, W., Coelho, L. P., Dolai, S., Fullam, A., Glazek, A., Hercog, R., Herrema, H., et al. (2022). Drivers and Determinants of Strain Dynamics Following Faecal Microbiota Transplantation. *Nature Medicine* 28(9), 1902–1912.

Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one* 6(3).

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nature Communications* 5(1), 1–12.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature Methods* 14(11), 1063–1071.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14), 2068–2069.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 9(8), 811–814.

Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell* 164(3), 337–340.

Sevim, V., Lee, J., Egan, R., Clum, A., Hundley, H., Lee, J., Everroad, R. C., Detweiler, A. M., Bebout, B. M., Pett-Ridge, J., et al. (2019). Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Scientific Data* 6(1), 1–9.

Shanahan, F., Ghosh, T. S., and O'Toole, P. W. (2021). The healthy microbiome—what is the definition of a healthy gut microbiome? *Gastroenterology* 160(2), 483–494.

Sharon, G., Cruz, N. J., Kang, D.-W., Gandal, M. J., Wang, B., Kim, Y.-M., Zink, E. M., Casey, C. P., Taylor, B. C., Lane, C. J., et al. (2019). Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. *Cell* 177(6), 1600–1618.

Shen, E. P. and Surawicz, C. M. (2008). Current treatment options for severe *Clostridium difficile*–associated disease. *Gastroenterology & Hepatology* 4(2), 134.

Shen, Y., Torchia, M. L. G., Lawson, G. W., Karp, C. L., Ashwell, J. D., and Mazmanian, S. K. (2012). Outer membrane vesicles of a human commensal mediate immune regulation and disease protection. *Cell Host & Microbe* 12(4), 509–520.

Shi, Z. J., Dimitrov, B., Zhao, C., Nayfach, S., and Pollard, K. S. (2022). Fast and accurate metagenotyping of the human gut microbiome with GT-Pro. *Nature Biotechnology* 40(4), 507–516.

Shkoporov, A. N., Clooney, A. G., Sutton, T. D., Ryan, F. J., Daly, K. M., Nolan, J. A., McDonnell, S. A., Khokhlova, E. V., Draper, L. A., Forde, A., et al. (2019). The human gut virome is highly diverse, stable, and individual specific. *Cell Host & Microbe* 26(4), 527–541.

Shkoporov, A. N., Khokhlova, E. V., Fitzgerald, C. B., Stockdale, S. R., Draper, L. A., Ross, R. P., and Hill, C. (2018). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nature Communications* 9(1), 1–8.

Shkoporov, A. N., Stockdale, S. R., Lavelle, A., Kondova, I., Heuston, C., Upadrasta, A., Khokhlova, E. V., Kamp, I. V. D., Ouwerling, B., Draper, L. A., Langermans, J. A. M., Ross, R. P., and Hill, C. (2022). Viral biogeography of the mammalian gut and parenchymal organs. *Nature Microbiology*.

Sibley, C. D., Grinwis, M. E., Field, T. R., Eshaghurshan, C. S., Faria, M. M., Dowd, S. E., Parkins, M. D., Rabin, H. R., and Surette, M. G. (2011). Culture enriched molecular profiling of the cystic fibrosis airway microbiome. *PloS one* 6(7), e22702.

Sieber, C. M., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., and Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* 3(7), 836–843.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* 22(3), 549–556.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15(2), 121.

Siranosian, B. A., Tamburini, F. B., Sherlock, G., and Bhatt, A. S. (2020). Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nature Communications* 11(1), 1–11.

Smillie, C. S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., Hohmann, E. L., Staley, C., Khoruts, A., Sadowsky, M. J., et al. (2018). Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host & Microbe* 23(2), 229–240.

Smith, B. J., Piceno, Y., Zydek, M., Zhang, B., Syriani, L. A., Terdiman, J. P., Kassam, Z., Ma, A., Lynch, S. V., Pollard, K. S., et al. (2022). Strain-resolved analysis in a randomized trial of antibiotic pretreatment and maintenance dose delivery mode with fecal microbiota transplant for ulcerative colitis. *Scientific Reports* 12(1), 1–14.

Smith, P. M., Howitt, M. R., Panikov, N., Michaud, M., Gallini, C. A., Bohlooly-y, M., Glickman, J. N., and Garrett, W. S. (2013). The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science* 341(6145), 569–573.

Song, J. H. and Kim, Y. S. (2019). Recurrent *Clostridium difficile* infection: risk factors, treatment, and prevention. *Gut and Liver* 13(1), 16.

Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K., and Shanks, O. C. (2017). Quantitative CrAssphage PCR assays for human fecal pollution measurement. *Environmental Science & Technology* 51(16), 9146–9154.

Staley, C., Halaweish, H., Graiziger, C., Hamilton, M. J., Kabage, A. J., Galdys, A. L., Vaughn, B. P., Vantanasiri, K., Suryanarayanan, R., Sadowsky, M. J., et al. (2021). Lower endoscopic delivery of freeze-dried intestinal microbiota results in more rapid and efficient engraftment than oral administration. *Scientific Reports* 11(1), 1–9.

Staley, C., Kaiser, T., Vaughn, B. P., Graiziger, C., Hamilton, M. J., Kabage, A. J., Khoruts, A., and Sadowsky, M. J. (2019). Durable long-term bacterial engraftment following encapsulated fecal microbiota transplantation to treat Clostridium difficile infection. *MBio* 10(4), e01586–19.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., and DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* 178(3), 591–599.

Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35(11), 1026–1028.

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of Bacteriology* 194(16), 4151–4160.

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* 10(12), 1196–1199.

Suskind, D. L., Brittnacher, M. J., Wahbeh, G., Shaffer, M. L., Hayden, H. S., Qin, X., Singh, N., Damman, C. J., Hager, K. R., Nielson, H., et al. (2015). Fecal microbial transplant effect on clinical outcomes and fecal microbiome in active Crohn's disease. *Inflammatory Bowel Diseases* 21(3), 556–563.

Sutton, T. D. and Hill, C. (2019). Gut bacteriophage: current understanding and challenges. *Frontiers in Endocrinology* 10, 784.

Szamosi, J. C., Forbes, J. D., Copeland, J. K., Knox, N. C., Shekarriz, S., Rossi, L., Graham, M., Bonner, C., Guttman, D. S., Van Domselaar, G., et al. (2020). Assessment of inter-laboratory variation in the characterization and analysis of the mucosal microbiota in Crohn's disease and ulcerative colitis. *Frontiers in Microbiology* 11, 2028.

Talley, N. J., Abreu, M. T., Achkar, J.-P., Bernstein, C. N., Dubinsky, M. C., Hanauer, S. B., Kane, S. V., Sandborn, W. J., Ullman, T. A., Moayyedi, P., et al. (2011). An evidence-based systematic review on medical therapies for inflammatory bowel disease. *Official journal of the American College of Gastroenterology| ACG* 106, S2–S25.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4(1), 1–14.

Temperton, B. and Giovannoni, S. J. (2012). Metagenomics: microbial diversity through a scratched lens. *Current Opinion in Microbiology* 15(5), 605–612.

Theriot, C. M., Bowman, A. A., and Young, V. B. (2016). Antibiotic-induced alterations of the gut microbiota alter secondary bile acid production and allow for *Clostridium difficile* spore germination and outgrowth in the large intestine. *mSphere* 1(1), e00045–15.

Tinsley, A., Williams, E., Liu, G., Elwir, S., Yoo, L., Melmed, G., and Sands, B. (2013). The incidence of influenza and influenza-related complications in inflammatory bowel disease patients across the United States: 1833. *Official journal of the American College of Gastroenterology| ACG* 108, S554.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., et al. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* 21(1), 1–21.

Torres, J., Mehandru, S., Colombel, J.-F., and Peyrin-Biroulet, L. (2017). Crohn's disease. *The Lancet* 389(10080), 1741–1755.

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011). Next generation sequence assembly with AMOS. *Current Protocols in Bioinformatics* 33(1), 11–8.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research* 27(4), 626–638.

Tsai, I. J., Otto, T. D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology* 11(4), 1–9.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449(7164), 804–810.

Van Dongen, S. and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. In: *Bacterial Molecular Networks*. Springer, 281–295.

Van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., Vos, W. M. de, Visser, C. E., Kuijper, E. J., Bartelsman, J. F., Tijssen, J. G., et al. (2013). Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *New England Journal of Medicine* 368(5), 407–415.

Vaughn, B. P., Vatanen, T., Allegretti, J. R., Bai, A., Xavier, R. J., Korzenik, J., Gevers, D., Ting, A., Robson, S. C., and Moss, A. C. (2016). Increased intestinal microbial diversity following fecal microbiota transplant for active Crohn's disease. *Inflammatory Bowel Diseases* 22(9), 2182–2190.

Vermeire, S., Joossens, M., Verbeke, K., Wang, J., Machiels, K., Sabino, J., Ferrante, M., Van Assche, G., Rutgeerts, P., and Raes, J. (2016). Donor species richness determines faecal microbiota transplantation success in inflammatory bowel disease. *Journal of Crohn's and Colitis* 10(4), 387–394.

Vich Vila, A., Imhann, F., Collij, V., Jankipersadsing, S. A., Gurry, T., Mujagic, Z., Kurilshikov, A., Bonder, M. J., Jiang, X., Tigchelaar, E. F., et al. (2018). Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Science Translational Medicine* 10(472), eaap8914.

Voigt, A. Y., Costea, P. I., Kultima, J. R., Li, S. S., Zeller, G., Sunagawa, S., and Bork, P. (2015). Temporal and technical variability of human gut metagenomes. *Genome Biology* 16(1), 1–12.

Walters, W. A., Xu, Z., and Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS letters* 588(22), 4223–4233.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73(16), 5261–5267.

Wang, S., Jiang, Y., and Li, S. (2021). PStrain: an iterative microbial strains profiling algorithm for shotgun metagenomic sequencing data. *Bioinformatics* 36(22-23), 5499–5506.

Watson, A. R., Fuessel, J., Veseli, I., DeLongchamp, J. Z., Silva, M., Trigodet, F., Lolans, K., Shaiber, A., Fogarty, E., Runde, J. M., et al. (2021). Adaptive ecological processes and metabolic independence drive microbial colonization and resilience in the human gut. *bioRxiv*.

Weingarden, A., Gonzalez, A., Vazquez-Baeza, Y., Weiss, S., Humphry, G., Berg-Lyons, D., Knights, D., Unno, T., Bobr, A., Kang, J., et al. (2015). Dynamic changes in short-and long-term bacterial composition following fecal microbiota transplantation for recurrent *Clostridium difficile* infection. *Microbiome* 3(1), 1–8.

Whelan, F. J., Verschoor, C. P., Stearns, J. C., Rossi, L., Luinstra, K., Loeb, M., Smieja, M., Johnstone, J., Surette, M. G., and Bowdish, D. M. (2014). The loss of topography in the microbial communities of the upper respiratory tract in the elderly. *Annals of the American Thoracic Society* 11(4), 513–521.

Whelan, F. J., Waddell, B., Syed, S. A., Shekarriz, S., Rabin, H. R., Parkins, M. D., and Surette, M. G. (2020). Culture-enriched metagenomic sequencing enables in-depth profiling of the cystic fibrosis lung microbiota. *Nature Microbiology*, 1–12.

Whipps, J., Lewis, K., and Cooke, R. (1988). Mycoparasitism and plant disease control. *Fungi in Biological Control Systems* 176(1), 161–187.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software* 4(43), 1686.

Willyard, C. (2018). New human gene tally reignites debate. *Nature* 558(7710), 354–356.

Wilson, B. C., Vatanen, T., Jayasinghe, T. N., Leong, K. S., Derraik, J. G., Albert, B. B., Chiavaroli, V., Svirskis, D. M., Beck, K. L., Conlon, C. A., et al. (2021). Strain engraftment competition and functional augmentation in a multi-donor fecal microbiota transplantation trial for obesity. *Microbiome* 9(1), 1–16.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences (USA)* 87(12), 4576–4579.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15(3), 1–12.

Wu, H.-J. and Wu, E. (2012). The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes* 3(1), 4–14.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2015). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32(4), 605–607.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32(4), 605–607.

Wu, Z., Greaves, J., Arp, L., Stone, D., and Bibby, K. (2020). Comparative fate of CrAssphage with culturable and molecular fecal pollution indicators during activated sludge wastewater treatment. *Environment International* 136, 105452.

Xie, F., Jin, W., Si, H., Yuan, Y., Tao, Y., Liu, J., Wang, X., Yang, C., Li, Q., Yan, X., et al. (2021). An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* 9(1), 1–20.

Yang, Z., Bu, C., Yuan, W., Shen, Z., Quan, Y., Wu, S., Zhu, C., and Wang, X. (2020). Fecal microbiota transplant via endoscopic delivering through small intestine and colon: no difference for Crohn's disease. *Digestive Diseases and Sciences* 65(1), 150–157.

Yap, C. X., Henders, A. K., Alvares, G. A., Wood, D. L., Krause, L., Tyson, G. W., Restuadi, R., Wallace, L., McLaren, T., Hansell, N. K., et al. (2021). Autism-related dietary preferences mediate autism-gut microbiome associations. *Cell* 184(24), 5916–5931.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486(7402), 222–227.

Youngster, I., Sauk, J., Pindar, C., Wilson, R. G., Kaplan, J. L., Smith, M. B., Alm, E. J., Gevers, D., Russell, G. H., and Hohmann, E. L. (2014). Fecal microbiota transplant for relapsing *Clostridium difficile* infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clinical Infectious Diseases* 58(11), 1515–1522.

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H., and Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 21(1), 1–15.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18(5), 821–829.

Zhang, F., Luo, W., Shi, Y., Fan, Z., and Ji, G. (2012). Should we standardize the 1,700-year-old fecal microbiota transplantation? *The American Journal of Gastroenterology* 107(11), 1755–author.

Zhang, Y., Bhosle, A., Bae, S., McIver, L. J., Pishchany, G., Accorsi, E. K., Thompson, K. N., Arze, C., Wang, Y., Subramanian, A., et al. (2022). Discovery of bioactive microbial gene products in inflammatory bowel disease. *Nature*, 1–7.

Zhu, F., Ju, Y., Wang, W., Wang, Q., Guo, R., Ma, Q., Sun, Q., Fan, Y., Xie, Y., Yang, Z., et al. (2020). Metagenome-wide association of gut microbiome features for schizophrenia. *Nature Communications* 11(1), 1–10.

Zivkovic, A. M., German, J. B., Lebrilla, C. B., and Mills, D. A. (2011). Human milk glycobiome and its impact on the infant gastrointestinal microbiota. *Proceedings of the National Academy of Sciences (USA)* 108(supplement_1), 4653–4658.

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology* 37(2), 179–185.