UNIVERSITY OF CALIFORNIA,
IRVINE

Investigating the interplay between diet, the human gut microbiota, and cancer

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Julio Avelar-Barragan

Dissertation Committee:
Associate Professor Katrine L. Whiteson, Chair
Professor Jennifer B.H. Martiny
Associate Professor Naomi S. Morrissette
Assistant Professor Elizabeth N. Bess

2023

DEDICATION

To

my parents

in recognition of your sacrifice and support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to express my gratitude for my principal advisor, Dr. Katrine Whiteson, for supporting my growth both as an individual and as a scientist. Your continued support, while also challenging me to do my best, has resulted in a fulfilling experience in graduate school. As my mentor, you treated me with humility, and you always encouraged me to be creative and to have fun. Outside of science, you also showed me what it takes to be a good leader. You were quick to deliver praise, and humble during the times of adversity. Again, thank you for taking me into your lab.

I am also incredibly grateful for my wonderful dissertation committee, including Dr. Jennifer Martiny, Dr. Elizabeth Bess, and Dr. Naomi Morrissette, who are some of the most intelligent individuals I have ever had the pleasure of working with. I appreciate everyone's thoughtful feedback, especially with my third dissertation chapter, and their continued contribution towards my success.

Next, I would like to extend a big thank you to all my friends that I have made during my time here at UCI. Thank you to all the previous Whiteson lab alumni, including Stephen Wandro, Tara Gallagher, Joann Phan, Andrew Oliver, and Whitney England, for receiving me with open arms into the lab and for all the wonderful memories. I would also like to thank my current lab members, Jason Rothman, Sage Dunham, Mirjam Zuend, Alisha Monsibais, and Eric Adams who are all brilliant scientists and even more brilliant friends. Outside of the lab, I would like to show my appreciation for the microbial group, especially Claudia Weihe and Elsa Abs, who have been some of my biggest supporters throughout graduate school. Lastly, I would like to thank Clark Hendrickson, Matthew Gargus, Julia De

Rogatis, Emily Neubert, Kristin Barbour, Alberto Barron, Cynthia Rodriguez, and Sarai Finks for some fantastic memories at UCI.

Outside of UCI, I would like to thank my close friends Ramon Guerrero, Rodrigo Aguayo, Bernabe Villalobos, Irais Cardenas, Puja Mazumder, and Rashika Choudhary for supporting me through the years. You all have been a blessing and I am eternally grateful for the good and bad times I had to share with everyone during graduate school. It has been such a pleasure to grow with you all. Next, I would like to show my appreciation for Proma Mazumder, who is an incredible friend and human being. You showed me that I have what it takes to make the world a better place. You are an excellent role model and inspiration. Thank you for being there for me in my darkest of times, and I will never forget the impact you have had on my life. Lastly, I would like to thank my amazing friend, Karla Viramontes. We had a late start to our friendship, but we have quickly made up the time with all the adventures we have been on. You are such a joy to be around, and you never fail to cheer me up. Thank you for pushing me to do my best, and for providing new perspective. Without a doubt, you are the reason my final year in graduate school has been so enjoyable. I love the confidence and energy you bring, and I eagerly await wherever life takes us next!

Additionally, I would like to express my appreciation for all the patients who made this research possible, some of whom are no longer with us. Some days, when the workload was heavy and I was stressed, I would find inspiration in the humility behind the data. Thank you for reminding me that our time and health is precious. Thank you for allowing me to learn and grow because of your contributions.

Above all, I would like to thank my parents, Jose Luis Avelar and Gabriela Avelar, whose sacrifices have allowed me to get to where I am today. They immigrated here from Mexico in their teens and have worked tirelessly since to give me and my brothers a future better than their own. Because of them, I never had to worry about food, clothes, or shelter and I could always focus on my studies. They achieved the American dream, and now, I am incredibly proud to be the first person in all my family to receive their doctorate. Never in my lifetime will I ever be able to repay you both for all that you have done for me, but at the very least I hope I have made you both proud. I love you mom and dad.

Chapter 1 of this thesis includes copyrighted material, originally appearing in *npj Biofilms and Microbiomes*. It is used with permission from Springer Nature. The co-authors listed in this publication are Lauren DeDecker, Zachary Lu, Bretton Coppedge, William Karnes, & Katrine Whiteson.

VITA

Julio Avelar-Barragan

A. Education

- University of California Riverside, Riverside, CA: B.S. (Microbiology, GPA: 3.87, *magna cum laude*) June 2017.
- University of California Irvine, Irvine, CA: Ph.D. (Molecular Biology & Biochemistry, GPA: 3.93) January 2023.

B. Honors and Awards

- Rose Hills Foundation Science & Engineering Fellowship, October 2022
- William F. Holcomb Scholarship, June 2022
- Graduate Assistance in Areas of National Need (GAANN) Fellowship, January 2022
- MB&B Outstanding Graduate Student Award, September 2021
- UCI Inclusive Excellence Ambassador Fellowship, June 2020
- UCI NIH-IMSD T32 Fellowship, January 2020
- Lake Arrowhead Microbial Genomics Best Poster Award, September 2018
- UCI NIH-IMSD T32 Fellowship, July 2017

C. Service

- Bioinformatics consultant, UCI Microbiome Center, UCI (July 2021-September 2021).
- Peer mentor, Office of Inclusive Excellence, UCI (July 2020-September 2020).
- Manuscript Peer-Reviewer, *Nature Communications*, reviewed with Katrine Whiteson, UCI (February 2020).
- Manuscript Peer-Reviewer, *Microbiology and Molecular Biology Reviews*, reviewed with Katrine Whiteson, UCI (October 2018).

D. Teaching Experience

- *Molecular Biology Laboratory*: Led the lab portion of the class as a teaching assistant, 15 students. UCI, Fall 2021.
- *General Microbiology*: Teaching assistant for a lecture-style class on microbial pathogens, 200-250 students per quarter. UCI, Fall 2020 and 2021.

- *Experimental Microbiology Laboratory:* Led the lab portion of the class as a teaching assistant, 20 students. UCI, Fall 2020.
- *Microbiome Data Analysis in R*: A workshop held by the UCI Microbiome Center on microbiome data analysis, 20-30 students per workshop. UCI, February 2018-March 2018

E. Outreach

- Guest lecturer for the UCI COSMOS program, July 2021. Lectured high school students interested in STEM about the human gut microbiome and dietary prebiotics.
- Journal club instructor for the Minorities in Science Program, June 2020-September 2020: A weekly journal club held by the UCI Initiative for Maximizing Student Development.
- Senior mentor for the ALMA science academy at James Madison Elementary, Santa Ana, October 2017-July 2018. Mentored and lectured underrepresented elementary school students to maximize diversity in STEM.

F. Publications

- Dunham, S. J., McNair, K. A., Adams, E. D., **Avelar-Barragan, J.**, Forner, S., Mapstone, M., & Whiteson, K. L. (2022). Longitudinal analysis of the gut microbiome in the 5xfAD mouse model of Alzheimer's disease. Accepted at mBio. https://doi.org/10.1101/2022.03.02.482725
- **Avelar-Barragan, J**., DeDecker, L., Lu, Z., Coppedge, B., Karnes, W., & Whiteson, K. (2022). Distinct Colon Mucosa Microbiomes associated with Tubular Adenomas and Serrated Polyps. NPJ Biofilms and Microbiomes. https://doi.org/10.1038/s41522-022-00328-6
- Oliver, A., El Alaoui, K., Haunschild, C., **Avelar-Barragan, J.**, Mendez Luque, L.F., Whiteson, K.L., Fleischman, A.G. (2021). Fecal microbial community composition in myeloproliferative neoplasm patients is associated with an inflammatory state. Microbiology Spectrum. https://doi.org/10.1128/spectrum.00032-22
- Jeney, S. E., **Avelar-Barragan, J**., Whiteson, K., Chang, J., Dutta, S., & Lane, F. (2021). Fecal Putative Uropathogen Abundance and Antibiotic Resistance Gene Carriage in Women With Refractory Recurrent Urinary Tract Infection Treated With Fecal Microbiota Transplantation. Female Pelvic Medicine & Reconstructive Surgery. doi: https://doi.org/10.1097/SPV.0000000000001090
- DeDecker, L., Coppedge, B., **Avelar-Barragan, J.**, Karnes, W., & Whiteson, K. (2021). Microbiome distinctions between the CRC carcinogenic pathways. Gut Microbes, 13(1), 1-12. https://doi.org/10.1080/19490976.2020.1854641

ABSTRACT OF THE DISSERTATION

Investigating the interplay between diet, the human gut microbiota, and cancer

by

Julio Avelar-Barragan

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2023

Professor Katrine L. Whiteson, Chair

As the world continues to industrialize, there has been a decreased incidence of infectious diseases and an increased incidence in non-communicable chronic diseases, such as obesity and cancer. One potential causal factor is the human microbiome, which refers to the collection of commensal bacteria, fungi, viruses, archaea, and other microorganisms that inhabit our bodies. The human microbiome has a collective genome which outnumbers human genes by 150-to-1, and it produces an expansive repertoire of metabolites which affect health. As such, the gut microbiome has major implications in digestion, educating the host immune system, preventing the colonization of pathogenic bacteria by occupying intestinal niches, and more.

Several components of the industrialized lifestyle have been known to alter the gut microbiome, leading to negative health consequences. The result is a microbiome that promotes inflammation and is associated with non-communicable chronic diseases. As these diseases become more prevalent in industrialized countries, there is an urgency to understand the role of the microbiota in human health. To investigate this issue, my research seeks to answer the following three questions: (1) What is the role of the

industrialized gut microbiome in the development of colorectal cancer? (2) How does diet affect the gut microbiome and inflammation in people with myeloproliferative neoplasms? (3) How is dietary fiber utilized by the gut microbiota of US and Moroccan individuals *in vitro*?

To answer the first question, samples were collected at the UCI Medical Center from 140 subjects during and after colonoscopy to characterize the microbiome associated with colorectal polyps. I used a combination of amplicon and shotgun metagenomic sequencing approaches to understand the effect of sampling method on microbial composition, and to describe the microbiome associated with healthy tissue and two types of colorectal polyps. I discovered that sampling method significantly explained 10-15% of the variation observed in microbiome composition. Additionally, using samples obtained from the colon mucosa, I was able to find associations with microbiome composition and colorectal polyps derived from the serrated pathway of colorectal carcinogenesis, such as a depletion in the lignan-degrading microbe, *Eggerthella lenta*. Lastly, I was able to use the microbiome to inform machine learning classifiers to accurately distinguish between healthy and polyp-bearing samples (Area under curve = 0.87-0.99).

To address the second question, we collaborated with Dr. Angela G. Fleischman, who conducted a dietary intervention in subjects afflicted with myeloproliferative neoplasms, a class of blood cancer. In this 15-week clinical trial, 28 individuals were assigned to receive dietary counseling following either a Mediterranean style eating pattern or one following U.S. guidelines. Blood and fecal samples were collected to examine inflammation and the gut microbiome, respectively. Using shotgun metagenomic sequencing, I discovered no significant alterations in gut microbiome diversity and composition due to a Mediterranean diet. I did find a

significant association between the gut microbiome and myeloproliferative neoplasm subtype, explaining approximately 6% of the variance in microbiome composition. Lastly, I found several significant correlations between microbial species, function, and cytokine concentrations.

My first two chapters suggested a link between the microbiome and dietary fiber, thus, in my final chapter I explored the effect of dietary fiber on microbial growth *in vitro*. In these experiments, I took the feces of 15 healthy individuals from the US and 15 Moroccans and cultured them anaerobically for 24 hours in the absence and presence of inulin, pectin, or psyllium husk. I found significant cohort effects on the microbiome, with US samples becoming dominated with *Bifidobacterium* and Moroccan samples becoming dominated with *Clostridia.* Furthermore, I demonstrate that pectin and psyllium husk perform differential enrichment of microbes and their associated carbohydrate-active enzymes.

INTRODUCTION

Industrialization has improved human health, and drastically extended the average life span. This is in part because modern medicine and improved sanitation have reduced the mortality, morbidity, and frequency of infectious disease. However, with the reduction of infectious disease, non-communicable chronic infectious disease incidence has disproportionately risen. This includes conditions like obesity, type 2 diabetes, autoimmune disorders, cancer, and cognitive diseases. One reason this might be occurring is the hygiene hypothesis, which states that the limited exposure to microbes, especially early in life, is causing deficiencies in immune regulation and function. [1] Indeed, the human body is inhabited with billions of commensal microorganisms, such as various bacteria, archaea, fungi, and viruses, which not only help educate our immune system, but also aid in digestion, out-compete pathogens, and more. [2-4] This collective of microorganisms is referred to as the human microbiome, and a significant proportion of metabolites in our bodies are produced by microbes. [5] Consequently, the microbiome has a substantial impact on human health.

Microbiome health is typically measured by alterations in diversity, composition, and function, and dysbiosis refers to an imbalance in the microbiome associated with disease. Conventionally, a more diverse gut microbiome is considered healthy because it provides functional redundancy among microbial community members, making the microbiome less susceptible to perturbations overall. [6] Microbial composition is another measure of gut microbiome health because there are microbes associated with beneficial metabolisms, like the fermentation of dietary fiber, as well as those associated with disease and the production of toxins or carcinogenic compounds. [7] Overall, diversity and

composition influence microbiome function, which can promote inflammation and disease if dysbiosis occurs. [1] Components of the industrialized lifestyle which negatively affect the gut microbiome includes antibiotic usage, physical inactivity, the consumption of processed foods, food high in fat and low in fiber, drug abuse, alcohol consumption, and chronic stress. [8-12]

One disease associated with a dysbiotic microbiome is colorectal cancer (CRC). It has been hypothesized that the increased incidence of CRC in industrialized countries is because of a high fat, low fiber diet. [13] For example, one mechanism is the metabolism of bile acids by the gut microbiome. In this scenario, primary bile acids are secreted by the host to digest excess fat. The bile acids then migrate to the large intestine, where they are converted into secondary bile acids by the gut microbiome. [14] Secondary bile acids have been demonstrated to promote inflammation and the development of CRC through the farsenoid X receptor. [14] The majority of CRCs develop through the adenoma-carcinoma sequence; however, a subset develops through an alternative mechanism called the serrated pathway. [15] The role of the gut microbiome has been described in the adenoma-carcinoma sequence, but its role is less clear in the serrated pathway. In my first chapter, I characterize the microbiome of samples collected during and after colonoscopy and examine the role of the industrialized gut microbiome in both pathways of CRC development. Since there is an over-reliance of fecal samples for capturing the microbiome in CRC research, I also described how mucosal samples collected directly from the colon compared to conventional fecal samples.

In my second chapter, I examined how diet affects the gut microbiome and inflammation in humans with myeloproliferative neoplasms (MPN), a type of blood cancer.

Unlike colorectal cancer, whose incidence is primarily driven by environmental factors, blood cancer development is influenced by host genetics. Specifically, myeloproliferative neoplasms are caused by mutations in the Philadelphia chromosome originating in the bone marrow. [16] There are three subtypes of Philadelphia-negative MPNs, but all share the overproduction of myeloid lineage cells. These cells give rise to red and white blood cells. The overproduction of white blood cells results in characteristically elevated systemic inflammation which drastically reduces the quality of life for those afflicted with the disease. MPN most frequently occurs in elderly individuals, and the disease often progresses slowly. Additionally, there are no definitive cures for MPN, other than risky bone marrow transplantations. [17] Thus, MPN treatment focuses on watchful waiting and symptom management. [17] Current pharmacological treatments are inadequate and carry serious side effects, therefore, there is a need to explore alternative therapies. Previous research has established alterations in the gut microbiome of MPN patients, and the microbiome can be manipulated with diet. [18, 19] The Mediterranean diet has been proven to reduce inflammation; therefore, my second chapter characterizes the gut microbiome of MPN patients during a Mediterranean diet intervention and its association with inflammation. [20]

One component believed to contribute to the reduced inflammation caused by the Mediterranean diet is its high fiber content. Fibers are polymeric carbohydrates that resist digestion from host enzymes. Fiber travels to the large intestine where is fermented by the gut microbiome using Carbohydrate Active Enzymes (CAZymes). Many of the resultant metabolites are beneficial for the host, such as short-chain fatty acids (SCFAs). One SCFA, butyrate, is the primary energy source of colonocytes. [21] Butyrate has also been

demonstrated to epigenetically regulate gene expression as a histone deacetylase and is anti-inflammatory. [21] Insufficient dietary fiber promotes the overabundance of mucus degrading microbes, decreases intestinal mucus production, and decreases the expression of tight junction proteins. [22] All of this can result in a condition called "leaky gut" in which gut microbes can cross the intestinal barrier, causing inflammation. [22] Chronic inflammation is the basis of several non-communicable diseases, such as CRC, inflammatory bowel disease, obesity, and other cognitive and metabolic diseases. [22] As such, there is substantial interest in leveraging prebiotic fibers to strategically manipulate the gut microbiome to promote health. Several studies with high fiber dietary interventions have been performed, but the results have been mixed. [23] Additionally, fibers have different monomers and chemical linkages which not all microbes can degrade. [24] It is not clear how different fibers are fermented by the gut microbiome, and if fiber can be used to enrich specific taxa. In my final chapter, I use three common dietary fibers to explore the relationship between fiber and the gut microbiome.

As research continues to reveal the impact of the microbiome, it is becoming increasingly clear that humans are holobionts and that our health is contingent on the many billions of microbes living on or within us. Rather than examining the host in isolation, more research is needed which incorporates the idea that humans are their own ecosystem. Together, my thesis seeks to elucidate how the food we eat affects the diversity, composition, and function of our gut microbiome, and how that is associated with cancer and inflammation.

REFERENCES

1. Sonnenburg, J. L. & Sonnenburg, E. D. Vulnerability of the industrialized microbiota. *Science* **366**, eaaw9255 (2019).
2. Ahn, J. *et al.* Human Gut Microbiome and Risk for Colorectal Cancer. *JNCI: Journal of the National Cancer Institute* **105**, 1907–1911 (2013).
3. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474**, 327–336 (2011).
4. Heintz-Buschart, A. & Wilmes, P. Human Gut Microbiome: Function Matters. *Trends in Microbiology* **26**, 563–574 (2018).
5. Dorrestein, P. C., Mazmanian, S. K. & Knight, R. Finding the Missing Links among Metabolites, Microbes, and the Host. *Immunity* **40**, 824–832 (2014).
6. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015).
7. Ulger Toprak, N. *et al.* A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clinical Microbiology and Infection* **12**, 782–786 (2006).
8. Moloney, R. D., Desbonnet, L., Clarke, G., Dinan, T. G. & Cryan, J. F. The microbiome: stress, health and disease. *Mamm Genome* **25**, 49–74 (2014).
9. Capurso, G. & Lahner, E. The interaction between smoking, alcohol and the gut microbiome. *Best Practice & Research Clinical Gastroenterology* **31**, 579–588 (2017).
10. Zinöcker, M. & Lindseth, I. The Western Diet–Microbiome-Host Interaction and Its Role in Metabolic Disease. *Nutrients* **10**, 365 (2018).

11. Mailing, L. J., Allen, J. M., Buford, T. W., Fields, C. J. & Woods, J. A. Exercise and the Gut Microbiome: A Review of the Evidence, Potential Mechanisms, and Implications for Human Health. *Exercise and Sport Sciences Reviews* **47**, 75–85 (2019).
12. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4554–4561 (2011).
13. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J Clin* **71**, 209–249 (2021).
14. Ocvirk, S. & O'Keefe, S. J. D. Dietary fat, bile acid metabolism and colorectal cancer. Seminars in Cancer Biology S1044579X2030208X (2020) doi:10.1016/j.semcancer.2020.10.003.
15. DeDecker, L., Coppedge, B., Avelar-Barragan, J., Karnes, W. & Whiteson, K. Microbiome distinctions between the CRC carcinogenic pathways. Gut Microbes 1–12 (2021) doi:10.1080/19490976.2020.1854641.
16. Fleischman, A. G. Inflammation as a Driver of Clonal Evolution in Myeloproliferative Neoplasm. *Mediators of Inflammation* **2015**, 1–6 (2015).
17. Spivak, J. L. Myeloproliferative Neoplasms. *N Engl J Med* **376**, 2168–2181 (2017).
18. Oliver, A. *et al.* Fecal Microbial Community Composition in Myeloproliferative Neoplasm Patients Is Associated with an Inflammatory State. *Microbiol Spectr* **10**, e00032-22 (2022).
19. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

20. Smidowicz, A. & Regula, J. Effect of Nutritional Status and Dietary Patterns on Human Serum C-Reactive Protein and Interleukin-6 Concentrations. *Advances in Nutrition* **6**, 738–747 (2015).
21. Hamer, H. M. *et al.* Review article: the role of butyrate on colonic function. *Alimentary Pharmacology & Therapeutics* **27**, 104–119 (2007).
22. Makki, K., Deehan, E. C., Walter, J. & Bäckhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host & Microbe* **23**, 705–715 (2018).
23. Myhrstad, M. C. W., Tunsjø, H., Charnock, C. & Telle-Hansen, V. H. Dietary Fiber, Gut Microbiota, and Metabolic Regulation—Current Status in Human Randomized Trials. *Nutrients* **12**, 859 (2020).
24. Cantu-Jungles, T. M. & Hamaker, B. R. New View on Dietary Fiber Selection for Predictable Shifts in Gut Microbiota. *mBio* **11**, e02179-19 (2020).

CHAPTER 1

Distinct Colon Mucosa Microbiomes associated with Tubular Adenomas and Serrated Polyps

AUTHORS: Julio Avelar-Barragan, Lauren DeDecker, Zachary N. Lu, Bretton Coppedge, William E. Karnes, Katrine L. Whiteson

DOI: https://doi.org/10.1038/s41522-022-00328-6

ABSTRACT

Colorectal cancer is the second most deadly and third most common cancer in the world. Its development is heterogenous, with multiple mechanisms of carcinogenesis. Two distinct mechanisms include the adenoma-carcinoma sequence and the serrated pathway. The gut microbiome has been identified as a key player in the adenoma-carcinoma sequence, but its role in serrated carcinogenesis is less clear. In this study, we characterized the gut microbiome of 140 polyp-free and polyp-bearing individuals using colon mucosa and fecal samples to determine if microbiome composition was associated with each of the two key pathways. We discovered significant differences between the microbiomes of colon mucosa and fecal samples, with sample type explaining 10–15% of the variation observed in the microbiome. Multiple mucosal brushings were collected from each individual to investigate whether the gut microbiome differed between polyp and healthy intestinal tissue, but no differences were found. Mucosal aspirate sampling revealed that the microbiomes of individuals with tubular adenomas and serrated polyps were significantly different from each other and polyp-free individuals, explaining 1–4% of the variance in the microbiome. Microbiome composition also enabled the accurate prediction of subject polyp types using Random Forest, which produced an area under

curve values of 0.87–0.99. By directly sampling the colon mucosa and distinguishing between the different developmental pathways of colorectal cancer, our study helps characterize potential mechanistic targets for serrated carcinogenesis. This research also provides insight into multiple microbiome sampling strategies by assessing each method's practicality and effect on microbial community composition.

INTRODUCTION

Colorectal cancer (CRC) is the second most deadly and third most common cancer globally, accounting for over 900,000 deaths in 2020.[1] The etiologies of CRC are multifactorial, with only 5-10% of cases being attributable to hereditary germline mutations.[2] Significant risk factors include diets high in red meat and low in fiber, obesity, physical inactivity, drug and alcohol usage, and chronic bowel inflammation.[3-6] Each of these factors is associated with compositional and functional changes in the collective community of bacteria, fungi, archaea, and viruses that inhabit the colon.[7-10] Commonly referred to as the gut microbiome, this community of microorganisms has been identified as a potential regulator of CRC initiation and progression.

Colorectal polyp formation precedes cancer development and is influenced by various environmental factors and host genetics. Polyps most commonly progress into malignancy through the adenoma-carcinoma sequence.[11] This pathway is characterized by chromosomal instability and mutations in the adenomatous polyposis coli (APC) gene, KRAS oncogene, and TP53 tumor suppressor gene.[12] Alternatively, 15 to 30% of CRCs develop through the serrated pathway.[13] This pathway is characterized by the epigenetic hypermethylation of gene promoters to produce a CpG island methylator phenotype.[13] In addition to the epigenetic inactivation of tumor suppressor genes, BRAF or KRAS mutations

are also common.[13] The serrated pathway often results in the production of hyperplastic polyps (HPPs), traditional serrated adenomas (TSAs), and sessile serrated polyps (SSPs).[14] Premalignant polyps from both pathways can be screened for and removed during colonoscopy to prevent CRC formation, but incomplete polyp resection or escaped detection can result in the development of interval cancers. Compared to other colorectal polyps, SSPs are disproportionately responsible for interval cancers, as their flat morphology makes them difficult to detect.[15] Additional detection methods, such as SSP-specific biomarkers, would assist with CRC prevention.

One potential avenue for polyp-specific biomarker discovery is the gut microbiota. SSPs often overexpress mucin forming genes, like MUC6, MUC5aC, MUC17, and MUC2, producing a mucus cap, which may harbor unique, mucin-degrading microbes.[16] Finding microbiome alterations in patients consistent with the presence of SSPs would enable gastroenterologists to personalize their technique and screening frequency for these higher risk patients. Additionally, elucidating the microbiome alterations specific to the adenoma-carcinoma sequence or the serrated pathway would help better understand the mechanisms of how particular microbes, their metabolites, and dysbiosis may contribute to colorectal carcinogenesis.

Studies comparing the microbiomes of these two pathways with healthy controls have yet to discover differences between healthy individuals and those with serrated polyps.[17-19] One reason for this may be the dominant use of stool for characterizing the microbiome, which does not accurately represent microbes adherent to the intestinal epithelium.[20,21] In this regard, we hypothesized that colon mucosa samples would more accurately reflect the composition of microbes intimately associated with colorectal polyps.

To investigate this, and the role of the microbiome in the adenoma-carcinoma and serrated pathways, we used multiple sampling techniques to obtain microbiome samples during and after colonoscopy from polyp-free individuals or those with tubular adenomas (TA), HPPs, or SSPs. When possible, mucosal brush samples from the same individual were collected from polyps and the healthy colon tissue opposite from these polyps. Stool samples were also collected 4-6 weeks after colonoscopy. We used a combination of amplicon (16S and ITS) and shotgun sequencing to study the microbial communities of samples. The purpose of our work was to 1) develop and compare microbiome sampling methods during colonoscopy; 2) characterize the microbiomes of polyp and healthy tissue samples within the same individuals; and 3) identify microbes or microbial genes specific to CRC precursors in the adenoma-carcinoma sequence versus the serrated pathway. Our key hypothesis was that there would be distinct differences between the microbiomes of individuals with tubular adenomas versus serrated polyps.

RESULTS

Description of the Study Cohort, Samples, and Data Collected:

We collected 1,883 mucosal brushes, mucosal aspirates, lavage aspirates, and fecal samples from 140 individuals with and without colorectal polyps (Supplementary Table 1.1). Of those, 50 individuals were polyp-free, 45 had at least one tubular adenoma, and 33 had at least one serrated polyp (Figure 1.1). The remaining 12 subjects had missing or unknown pathologies. We generated data from two sample sets. The first sample set was characterized using 16S and ITS sequencing, while the second sample set was analyzed using shotgun sequencing. Details on the number of samples, sample types, polyp types,

and subject characteristics for each dataset can be found in Table 1.1 and Supplementary Tables 1.2-1.4.

Microbiomes of Mucosal and Lavage Samples are similar to each other but different from those in Feces:

Our first objective was to determine whether microbiome composition varied between sample types. We began by sequencing DNA from mucosal brushes, mucosal aspirates, and lavage aspirates from a subset of 38 individuals using 16S amplicon sequencing. Fecal samples were not included because they were collected later.

We observed no significant differences in Shannon diversity or richness across mucosal brushes, mucosal aspirates, and lavage aspirates (Linear mixed effects model, LME: $p > 0.05$, Figure 1.2A). Permutational multivariate analysis of variance (PERMANOVA) on Bray-Curtis dissimilarities revealed that the individual explained the greatest amount of variation in microbiome composition (PERMANOVA: $p = 0.001$, $R^2 = 0.51$; Supplementary Table 1.5). This analysis found no significant differences in the microbiomes associated with mucosal brushes, mucosal aspirates, and lavage aspirates from within the same individual (PERMANOVA: $p = 0.99$, $R^2 = 0.15$; Supplementary Table 1.5). The lack of significance was consistent with no discernable clusters based on sample type (Figure 1.2B). The abundances of three amplicon sequence variants (ASV) significantly differed across the three sampling methods – one from the *Gemellaceae* family and two *Streptococcus spp.* Abundances of these ASVs were higher in mucosal aspirates compared to mucosal brushes (ANCOM2: p-adj < 0.05; Supplementary Figure 1.1).

ITS2 sequencing was also performed on the same subset of samples to investigate the effect of sampling method on the fungal microbiome. We observed no differences in

Shannon diversity or richness across mucosal brushes, mucosal aspirates, and lavage aspirates (LME: $p > 0.05$, Figure 1.2C). Beta-diversity ordination by sample type demonstrated no discernable clustering (Figure 1.2D). Like 16S amplicon data, PERMANOVA analysis of Bray-Curtis dissimilarities showed that the individual significantly explained the greatest amount of variation in fungal community composition (PERMANOVA: $p = 0.003$, $R^2 = 0.28$), with no significant associations between fungal community composition and our three sampling methods (PERMANOVA: $p = 0.36$, $R^2 = 0.38$; Supplementary Table 1.6).

Following the collection of fecal samples, we performed shotgun sequencing on a second subset of samples. Mucosal brushes were excluded from the second sample set because a pilot shotgun sequencing run revealed these samples contained a large proportion of human-derived reads (Supplementary Figure 1.2). Based on estimates of Shannon diversity and species richness, the microbiomes in fecal samples were significantly more diverse than those in the mucosal aspirates (LME: $p = 0.007$ and $p = 0.002$, respectively) and marginally more diverse than those in lavage aspirates (LME: $p = 0.053$ and $p = 0.047$, respectively; Figure 1.2E). Visualization of sample beta-diversities revealed a cluster of fecal samples that partially overlapped with mucosal and lavage aspirates (Figure 1.2F). PERMANOVA showed that the individual explained the greatest amount of variation in microbiome composition (PERMANOVA: $p = 0.001$, $R^2 = 0.72$; Supplementary Table 1.7). In comparison, sampling method explained 15% of variation in the microbiome (PERMANOVA: $p = 0.001$). Fecal samples had a mean relative abundance of 63% for Firmicutes, 27% for Bacteroides, 3.5% for Actinobacteria, and 4.5% for Proteobacteria. Mucosal aspirates and lavage aspirates were more similar and had a mean

relative abundance of 73% and 75% for Firmicutes, 15% and 11% for Bacteroides, 4.7% and 5.2% for Actinobacteria, and 4.0% and 6.6% for Proteobacteria, respectively (Supplementary Figure 1.3). Differential abundance analysis revealed 42 microbes whose abundances significantly differed between fecal samples and mucosal aspirates (ANCOM2: p-adj < 0.05; Supplementary Table 1.8). Five microbes were differentially abundant between fecal samples and lavage aspirates (Supplementary Table 1.9), and no microbes were significantly different between mucosal aspirates and lavage aspirates (ANCOM2; p-adj > 0.05).

The Microbiomes of Polyps and Healthy Opposite Wall Tissue are similar within Individuals:

To characterize the microenvironment of polyps, 14 mucosal brush samples from 6 subjects were collected from polyps and healthy opposite wall tissue and sequenced as part of the first sample set (Figure 1.3A). Based on 16S sequencing, we observed no significant differences in Shannon diversity or richness between polyp and healthy opposite wall tissue from within the same individual (Figure 1.3B). We did observe significantly increased richness in samples from the left sided colon when compared to the right sided colon (Figure 1.3B, LME: p = 0.01). With respect to beta-diversity, there were no significant differences across polyp and healthy opposite wall tissue pairs (PERMANOVA: p = 0.87, R^2 = 0.18; Figure 1.3C; Supplementary Table 1.10). We were unable to identify any differentially abundant microbes between polyp and opposite wall tissue brushes. Microbiomes were mostly individualistic, with subject origin explaining 55% of the variance in microbiome composition (PERMANOVA: p = 0.02; Figure 1.3D; Supplementary Table 1.10).

Tubular Adenoma-bearing, Serrated Polyp-bearing, and Polyp-free Individuals have distinct Microbiomes:

We next reanalyzed all samples from the first and second sample sets to examine whether the subject's polyp type of a sample (polyp-free vs. tubular adenoma-bearing vs. serrated polyp-bearing) was significantly associated with microbial diversity and composition. In both 16S and shotgun data, we observed no significant differences between subject types based on Shannon diversity or richness estimates (LME: $p > 0.05$; Supplementary Figure 1.4). In ITS data, we observed significantly increased Shannon diversity, but not richness, in samples from polyp-free individuals when compared to those from TA-bearing individuals (LME: $p = 0.03$; Supplementary Figure 1.4). Beta diversity analysis of 16S and ITS data from the first sample set demonstrated that subject type explained 4% and 2% of the variance associated with the microbiome, respectively (16S PERMANOVA: $p = 0.001$; Supplementary Table 1.5 and ITS PERMANOVA: $p = 0.204$; Supplementary Table 1.6).

In the second sample set, we found significant associations between the microbiome and subject type explaining 2% of the observed variance (PERMANOVA: $p = 0.001$; Supplementary Table 1.7). This association was examined further by testing each pairwise subject type comparison within each sample type. In TA vs. SP-bearing mucosal aspirates, subject type significantly explained 2.7% of the variance associated with the microbiome (PERMANOVA: $p = 0.001$; Supplementary Table 1.11). The proportion of significant variance associated with subject type was reduced to 1.9% for polyp-free vs. TA-bearing mucosal aspirates (PERMANOVA: $p = 0.001$) and 1.5% for polyp-free vs. SP-bearing mucosal aspirates (PERMANOVA: $p = 0.001$; Supplementary Table 1.11). An association

between microbiome composition and subject type was not observed when testing lavage aspirates (PERMANOVA: p = 0.47; Supplementary Table 1.12) or fecal samples (PERMANOVA: p = 0.10; Supplement Table 1.13) alone.

We then performed an in-depth investigation of each subject type's microbiome using only second sample set mucosal aspirates, due to their larger comparable sample size. Differential abundance analysis demonstrated that *Eggerthella lenta* was significantly depleted in SP-bearing aspirates when compared to polyp-free ones (Kruskal-Wallis, KW: p-adj = 0.032). *E. lenta* also demonstrated a lower abundance in SP mucosal aspirates when compared to TA aspirates but this decrease was not significant (KW: p-adj = 0.099). Supplementary Figure 1.5 suggest that *E. lenta* was also depleted in 16S mucosal aspirates, but this result was not statistically significant either.

Despite few differentially abundant microbes, taxonomic visualization suggested that TA-bearing mucosal aspirates were distinct compared to polyp-free and SP-bearing mucosal aspirates (Figure 1.4A, Supplementary Figure 1.6). Therefore, we examined if microbial composition could be used to predict the subject type origin of mucosal aspirates. Random Forest (RF) classified mucosal aspirates from each pairwise subject type comparison with moderate to high accuracy, producing area under curve (AUC) values of 0.87 – 0.99 (Figure 1.4B). The top variables of importance for the classification of polyp-free versus TA-bearing mucosal aspirates were *Ruthenibacterium* sp., *Ruminococcus gnavus*, *Ruminococcus* sp., *Dorea* sp., and *Blautia* sp. (Figure 1.4C). For polyp-free versus SP-bearing RF classification, *Anaerostipes hadrus*, *Dorea longicatena*, *E. lenta, Clostridium ramosum,* and *Alistipes finegoldii* were the most important variables (Figure 1.4D). Lastly, *Gemmiger formicilis*, *E. lenta, Bifidobacterium* sp., *Ruthenibacterium* sp., and UBA7182

HGM12585 were the top microbes of importance for the SP-bearing versus TA-bearing RF classification (Figure 1.4E). Figure 1.4F displays the relative abundances for the top variables of importance in all RF comparisons.

Microbiome Functional Potential is distinct across Sampling Methods and Subject Types:

The functional characteristics of our shotgun metagenomes were next explored. Pathway analysis resulted in the discovery of 507 metabolic pathways, which were generally conserved across subject types (Figure 1.5A and Supplementary Figure 1.7). As a result, we did not identify any differentially abundant pathways (KW: p-adj > 0.05). Additionally, pairwise RF classification of functional pathways resulted in lower AUC values when compared to taxonomic RF classification (Supplementary Figures 1.8 and 1.9). Subsequently, we analyzed individual microbial genes, whose composition exhibited a higher correlation to microbial taxonomy (Mantel: p = 0.001, r = 0.70) when compared to functional pathways (Mantel: p = 0.001, r = 0.33). Like previous taxonomic results, we found that fecal samples had significantly increased Shannon diversity (LME: p = 0.034) and gene richness estimates (LME: p = 0.021) when compared to mucosal aspirates, but not mucosal lavages (Figure 1.5B). Principal coordinate analysis resulted in fecal samples clustering together, with no obvious clustering based on subject type (Figure 1.5C). This was supported by PERMANOVA, which confirmed an association between functional metagenome and sampling method, explaining 10.8% of the observed variance (PERMANOVA: p = 0.001; Supplementary Table 1.14). By comparison, the individual of origin explained approximately 76% of the observed variance in the functional microbiome

(PERMANOVA: p = 0.001; Supplementary Table 1.14) and subject type explained 1.3% of the observed variance (PERMANOVA: p = 0.001; Supplementary Table 1.14).

We concluded our analysis by searching for differentially abundant genes among subject types using mucosal aspirates, but did not find any after adjusting for the false discovery rate (KW: p-adj > 0.05). Supplementary Figure 1.10 demonstrates that the majority of the genes determined to be differentially abundant before FDR correction originated from the class *Coriobacteriia*, which *E. lenta* belongs to. Given that *E. lenta* metabolizes plant lignans in the gut and was found to be depleted in SP-bearing mucosal aspirates, we decided to examine which *E. lenta* specific carbohydrate active enzymes (CAZymes) were present in our metagenomes. We found six CAZymes, of which four had decreased abundance in SP-mucosal aspirates. These were a carbohydrate esterase, family 2 (CT2) and three glycosyl transferases from families 2, 28, and 51 (GT2, GT28, and GT51; Figure 1.5D). A complete list of differentially abundant genes, their functions, and taxonomy before FDR correction can be found in the supplement (Supplementary File 1).

DISCUSSION

In this study we used direct and indirect methods to sample the colon and characterize the microbiomes of polyp-free and polyp-bearing individuals. Using amplicon sequencing, we found that microbiomes of mucosal brushes and mucosal aspirates did not significantly differ in diversity or composition. In contrast, the microbiomes of fecal samples were significantly more diverse and compositionally distinct when compared to those from mucosal aspirates.

Due to their ease of collection, fecal samples are frequently used to study the human microbiome in the context of CRC. However, fecal samples poorly represent the microbiota

adherent to the colon mucosa, and instead capture those found in the intestinal lumen.[20,21] Their increased diversity and paucity of mucosa-associated microbes suggests that fecal samples are less ideal for studying premalignant polyps, which have fewer pronounced signatures of microbial dysbiosis when compared to carcinomas. This is supported by Peters *et al.*, who found greater compositional changes in the microbiomes of fecal samples from advanced conventional adenomas when compared to those from non-advanced adenomas.[17] The decreased sensitivity of fecal samples to detect CRC-associated microbes was also highlighted by their results demonstrating significant associations between the gut microbiome and distal conventional adenoma cases, but not proximal.[17] This is also likely why Peters *et al.* did not observe substantial differences in the microbial compositions of HPP, SSP, and healthy samples, as serrated polyps predominantly develop in the proximal colon.

Here, we report significant associations between the gut microbiome and mucosal aspirates obtained from both the proximal and distal colon. We also observed significant differences when comparing the microbiomes of polyp-free samples to SP-bearing ones using mucosal aspirates. No such differences were seen in fecal samples, but this result may be driven by a smaller sample size. Nevertheless, these data suggest that mucosal samples are sensitive enough to study the microbiome of colorectal polyps found within the proximal colon. This contradicts a study published by Yoon *et al.*, who found no significant compositional differences among the mucosa-associated gut microbiomes of polyp-free, TA, SSP, and CRC-bearing individuals.[18] The authors note, however, that their result was likely influenced by a small sample size, with only 6 samples per group and 24 samples total.

Compared to mucosal brushes, mucosal aspirates had a lower risk of damaging the intestinal epithelium, provided larger collection volumes for downstream sample processing, and resulted in lower proportions of human derived reads during shotgun sequencing. Both methods also had similar microbiome profiles. One caveat of our approach, however, is that we did not collect mucosal aspirates from polyp tissue directly, only from healthy tissue near polyps. Therefore, it is unclear whether the three differentially abundant microbes observed between mucosal brushes and aspirates was due to the sampling method used or the tissue site (Supplementary Figure 1). Certainly, more research is needed to further evaluate each sampling method, but we believe the advantages of mucosal aspiration outweigh the risk of mucosal brushing and any minor discrepancies in microbiome diversity and composition.

With respect to characterizing the hyperlocal microbiome of polyps and opposite colon wall tissue, mucosal brushing revealed no differences. One factor which could have disrupted any potential hyperlocal differences in the gut microbiota is the colonoscopy preparation and lavage. As part of the preparation, individuals were advised to adhere to a low fiber, clear liquid diet 24 hours prior to colonoscopy. Dietary fiber is important for maintaining the longitudinal and lateral organization of the microbiota within the colon, as giving mice a low fiber diet has been shown to disrupt the microbial organization their guts.[20] Additionally, changes in diet can rapidly shift the composition of the gut microbiome, often within 24 hours.[7,22,23] Another factor which could have potentially obscured the hyperlocal organization of colon epithelium further was the mechanical displacement caused by the laxative-based cleansing and colonoscopy rinse. Nevertheless, significant compositional differences between the microbiomes of samples taken from the

proximal and distal colon were observed, suggesting that broad microbial organization remained present in the gut after colonoscopy preparation and lavage. It is important to note that these claims are based on data from 14 samples from 6 individuals, therefore, additional studies with more samples are needed to validate the reproducibility of our findings.

Comparatively, compositional differences were observed in the gut microbiome across TA-bearing, SP-bearing, and polyp-free individuals using mucosal sampling. Notably, we demonstrated that the microbial composition of each subject type was distinct enough to accurately predict the origin of mucosal aspirates using RF. These findings suggest that the gut microbiome plays different roles in the adenoma-carcinoma sequence and the serrated pathway. In the adenoma-carcinoma sequence, the gut microbiome exists in, and potentially contributes to, an inflammatory environment to promote colorectal carcinogenesis.

Data obtained from second set mucosal aspirates supports that TA-bearing subjects had an altered microbiome composition associated with inflammation and CRC development. These samples trended towards a higher abundance of *Lachnospiraceae*, such as *Ruminococcus gnavus,* which has been previously associated with CRC and inflammatory bowel disease, and *C. scindens,* which can metabolize excess primary bile acids not absorbed by the small intestine into secondary bile acids (Supplementary Figure 1.6).[24-26] High concentrations of secondary bile acids can cause host oxidative stress, nitrosative stress, DNA damage, apoptosis, and mutations.[27] Secondary bile acids also act as farnesoid X receptor antagonists, resulting in enhanced *wnt* signaling in the adenoma-carcinoma sequence.[28] RF classification also identified *Bacteroides fragilis* as a top variable of

importance, which was elevated in TA mucosal aspirates. *B. fragilis* produces a metalloprotease that causes oxidative DNA damage and cleaves the tumor suppressor protein, E-cadherin.[29-31]

Unlike the adenoma-carcinoma sequence, the microbiome in the serrated pathway remains understudied. *Fusobacterium nucleatum,* which has been implicated in the adenoma-carcinoma sequence because of its ability to activate *wnt* signaling, has also been described as having a role in serrated CRC development.[32] *F. nucleatum* abundance is associated with serrated pathway lesions and features, such as mismatch repair deficiency, MLH1 methylation, CpG island methylator phenotype, and high microsatellite instability.[14] Here, we did not find differences in *F. nucleatum* abundances across HPPs, SSPs, Tas, or polyp-free controls. Instead, we most prominently found that *E. Lenta* and its CAZymes were depleted in mucosal aspirates from SP-bearing individuals, a result that spanned 16S and shotgun data.

E. lenta metabolizes inert plant lignans in the gut into bioactive enterolignans, such as enterolactone and enterodiol.[33] These enterolignans have anti-proliferative and anti-inflammatory effects, and help modulate estrogen signaling, lipid metabolism, and bile acid regulation.[34] They have also been associated with reduced cancer risk.[35] Diets rich in plant fiber have been associated with decreased CRC risk.[6,36] Fiber is fermented by the intestinal microbiota to produce short chain fatty acids, including acetate, butyrate, and propionate. Butyrate is the primary energy source for colonocytes and has anti-inflammatory and anti-tumor properties.[37-39] Butyrate also is involved in the epigenetic expression of genes as a histone deacetylase inhibitor.[40] In the serrated pathway, the gene SLC5A8, which mediates short chain fatty acid uptake into colonic epithelial cells, is frequently inhibited via

promoter methylation, suggesting that dietary fiber may be required for proper cellular epigenetic regulation.[41]

Further evidence of dietary fiber potentially playing a role in the serrated pathway was the identification of *A. hadrus* as the most important variable in differentiating polyp-free vs SP-bearing mucosal aspirates by RF. *A. hadrus* is a butyrate producing microbe and was depleted in SP-bearing mucosal aspirates. Taken together, we hypothesize that low dietary fiber consumption facilitates aberrant epigenetic modifications within colonocytes to promote serrated polyp development, but studies which combine both mucosal sampling methods and dietary information are needed to test this hypothesis.

In conclusion, the complex and individualistic nature of the human gut microbiome has made it difficult to mechanistically link the microbiome with colorectal carcinogenesis. By describing the association between the gut microbiota and two colorectal polyp types with several sampling methods, our study provides insight into potential mechanisms for the epigenetic-based serrated pathway of CRC. In addition, our data underscores the importance of distinguishing between different pathways of colorectal carcinogenesis when investigating the gut microbiome. Finally, transitioning future microbiome studies to mucosal sampling methods may enable the discovery of previously unassociated CRC microbes.

METHODS

Subject Recruitment and Criteria:

Individuals who presented for colonoscopy with indications of screening for, or a prior history of, colorectal polyps were asked to participate in the study. Written and informed consent was obtained from each subject and was required for participation.

Subjects who were pregnant, had taken antibiotics within 6 weeks of colonoscopy, or with known inflammatory bowel diseases, were excluded. In total, 140 individuals were recruited for this study. Of the 140 individuals, 50 were found to be polyp-free, 45 had one or more Tas, 33 had polyps originating from the serrated pathway (HPPs or SSPs), and 12 had unknown or other pathologies.

Colonoscopy Preparation, Procedure, and Sample Collection:

Before colonoscopy, subjects were asked to adhere to a clear liquid diet for 24 hours. Bowel cleansing was done using Miralax, or polyethylene glycol with electrolytes administered as a split dose, 12 and 5 hours before the procedure. Sample collection focused on two direct and two indirect microbiome sampling methods (Figure 1.1). The first direct sampling method involved brushing the mucosa of colon during colonoscopy (Method #1 in Figure 1.1). Brushing was performed on suspected polyps and on opposing healthy colon tissue to compare their microenvironments. Since mucosal brushes can potentially damage or agitate the intestine, we also employed a method of direct microbiome sampling in which colonoscopy washing fluid was sprayed directly on to the target mucosa and immediately re-suctioned into a storage vial (Method #2 in Figure 1.1). Participants with suspected polyps had mucosal washing aspirates taken on healthy tissue near the polyp, but no mucosal aspirates were taken from polyps directly. The first indirect sampling method involved collecting an aspirate of the post-colonoscopy lavage fluid (Method #3 in Figure 1.1). This lavage fluid was produced from rinsing the wall of the colon throughout the procedure and was collected in a container outside the subject. All samples were collected in sterile cryogenic tubes and placed on ice until the colonoscopy procedure was finished. Afterwards, the samples were stored at -80°C. Additional information

collected included indication for procedure, age, sex, ethnicity, BMI, family history, and findings, including the size, shape, location, and pathology of all polyps sampled.

Patient-directed Collection of Fecal Samples:

For the second indirect microbiome sampling method, subjects were encouraged to send follow-up fecal samples four to six weeks post-colonoscopy (Method #4 in Figure 1.1). Subjects were provided with a fecal collection kit, which contained collection equipment, prepaid shipping labels, and Zymo DNA/RNA shield preservation buffer (R1101). Subjects who complied were compensated $20 USD. Samples were returned via the United States Postal Service. After arrival, samples were stored at -80°C. Thirty-eight fecal samples were returned, bringing our total number of samples collected to 1,883. A summary of the sample types can be found in Supplementary Table 1.1.

Polyp and Subject Type Classification:

Polyp biopsies collected during colonoscopy were sent to a pathologist for classification. This information was then recorded for the corresponding mucosal brush and aspirate samples. Pathology reports were also used to broadly categorize all samples collected from an individual by their polyp pathology. We referred to this as the 'subject type' and the three categories were polyp-free subjects, tubular adenoma-bearing subjects (TA-bearing), and serrated polyp-bearing (SP-bearing) subjects, which included both HPPs and SSPs. For example, if a sample was taken from healthy intestinal tissue of an individual who was found to have a TA, that sample and all others from the same individual would be included in the TA-bearing subject type. Three individuals had both a TA and an SSP and were classified as SP-bearing subjects.

DNA Extraction:

Two separate DNA extractions were performed in this study, yielding two different sample sets (Table 1.1). Sample set 1 DNA extractions included mucosal brushes, mucosal aspirates, and lavage aspirates only. Sample set 2 DNA extractions occurred later and included mucosal aspirates, lavage aspirates, and fecal samples. All samples were thawed on ice for DNA extraction. For mucosal aspirates and lavage aspirate samples, 250 uL of fluid were taken from each sample and then DNA was extracted using ZymoBiomics DNA Miniprep Kit (D4300) according to the manufacturer's protocol. For mucosal brushes, 750 uL of ZymoBIOMICS Lysis Solution was mixed with the brushes in their original sterile cryogenic tube and vortexed for 5 minutes to suspend the contents of the brush into solution. The solution was then transferred and extracted according to the manufacturer's protocol. Fecal samples stored in Zymo DNA/RNA shield were thawed, mixed by vortexing, and 750 uL of the fecal plus buffer mix was extracted according to the manufacturer's protocol.

16S Amplicon Library Preparation and Sequencing:

Samples from the first set underwent 16S and ITS amplicon sequencing. We targeted the V4 region of the bacterial 16S rRNA gene using the 515F and 926R primers. For each sample, the V4 region was amplified using 25 uL polymerase chain reaction (PCR) volumes with the following reagents: 12.5 uL of 1x AccustartII PCR tough mix (QuantaBio 95142), 9.5 uL of PCR grade water, 1 uL of 10 mg/mL BSA, 0.5 ng of extracted genomic DNA, and 0.5 uL of 0.2 uM 515F and 926R primers each. The 515F primer contained the Illumina adapter sequence and barcode. Each sample was amplified using a thermocycler for 30 cycles (94°C for 3 min; 94°C for 45 sec, 55°C for 30 sec, 72°C for 20 sec; repeat steps 2-4 30 times; 72°C for 10 min). The resultant amplicons were quantified using the Qubit

dsDNA HS Assay Kit (Life technologies Q32851) according to the manufacturer's protocol and pooled at equimolar concentrations. The pooled amplicon library was cleaned and concentrated using Agencourt AMPure XP beads (Beckman-Coulter A63880) according to the manufacturer's protocol. Equimolar PhiX was added at 10% final volume to the amplicon library and sequenced on the Illumina MiSeq platform, yielding 300bp paired-end sequences. A total of 200 samples with an average of 41,578 +/- 35,920 (σ) reads per sample were obtained for 16S amplicons.

ITS Amplicon Library Preparation and Sequencing:

Fungi from the first sample set were characterized by targeting the ITS2 region of the 18S rRNA gene for amplification. We used the ITS9f and ITS4r primers, as described by Looby et al.[43] PCR was performed in 25 uL volumes, consisting of: 12.5 uL of 1x AccustartII PCR tough mix, 9.5 uL of PCR grade water, 1 uL of 10 mg/mL BSA, 0.5 ng of extracted genomic DNA, and 0.5 uL of 0.3 uM ITS9f and barcoded ITS4r primers each. Amplification was performed with the following thermocycler settings: 94°C for 5 min, 35 cycles of 95°C for 45 sec, 50°C for 1 min, 72°C for 90 sec, and a final extension step of 72°C for 10 min. Afterwards, we quantified, pooled, and cleaned our ITS2 amplicons using the same methods as our 16S amplicons. Our ITS2 library was combined with our 16S library and sequenced simultaneously in the reverse complementary orientation. This yielded 150 samples with an average of 22,252 +/- 17,000 (σ) ITS reads per sample.

Shotgun Library Preparation and Sequencing:

The second sample set was sequenced using shotgun sequencing. Libraries were prepared using the Illumina DNA prep kit (20018705), using our low-volume protocol.[44] Briefly, a maximum of 5 uL or 50 ng (whichever was reached first) of DNA from each

sample was tagmented with 2 uL of tagmentation master mix for 15 min at 55°C. Afterwards, 1 uL of tagmentation stop buffer was added to each sample and incubated at 37°C for 15 min. The samples were washed with the provided buffer according to the manufacturer's protocol, then PCR was performed with 12.5 uL reaction volumes with the following reagents: 6.25 uL of KAPA HiFi HotStart ReadyMix (Roche Life Science KK2602), 2.75 uL of PCR grade water, 1.25 uL of 1 uM i5 and i7 index adaptors each, and 0.5 uL of 10 uM forward and reverse KAPA HiFi polymerase primers each. PCR amplification was done with the settings: 72°C for 3 min, 98°C for 3 min, 12 cycles of 98°C for 45 sec, 62°C for 30 sec, 72°C for 2 min, and a final extension step of 72°C for 1 min. Samples were pooled and size selection was performed per the manufacturer's protocol. Libraries were packaged on dry ice and shipped overnight to Novogene Corporation Inc. (Sacramento, CA) to be sequenced using Illumina's Hiseq 4000 for 150 bp paired-end sequencing. This yielded 257 samples with an average of 1,267,359 +/- 690,384 (σ) reads per sample.

Taxonomic Assignment of Sequencing Data:

For first sample set, 16S and ITS amplicon sequences were processed using Qiime2-2019.1.[45] Demultiplexing was performed using the 'q2-demux' function with the 'emp-paired' preset. Sequencing reads were quality filtered, had chimeric sequences, PhiX, and singletons removed, and were clustered into amplicon sequence variants (ASVs) using the 'q2-dada2' function with the default parameters plus trunc_len_f = 280, trunc_len_r = 220, trim_left_f = 5, and trim_left_r = 5.[46] This reduced the average number of reads per sample to 30,051 +/- 24,768 (σ) for 16S amplicons, and 3,517 +/- 9,154 (σ) for ITS amplicons. Taxonomic assignment of 16S and ITS reads was done using the 'classify-sklearn' function in Qiime2 with the default parameters. The databases used for classification were the

Greengenes database (Version 13.8) for 16S data, and the UNITE database (Version 8.0) for ITS data.[47,48] This produced 182 samples with an average of 28,343 +/- 23,150 (σ) high-quality, taxonomically assigned reads per sample for 16S amplicons, and 131 samples with an average of 3,461 +/- 8,357 (σ) high-quality, taxonomically assigned reads per sample for ITS amplicons.

For second sample set shotgun data, we first removed sequencing adapters using the 'bbduk.sh' script from BBMap v38.79 with the default parameters.[49] Next, we demultiplexed our samples using 'demuxbyname.sh' script from BBMap using the default parameters. After demultiplexing, sequences were quality filtered using PRINSEQ++ v1.2 with the parameters trim_left = 5, trim_right = 5, min_len = 100, trim_qual_right = 28, and min_qual_mean = 25.[50] This yielded an average of 1,209,001 +/- 643,544 (σ) high quality reads per sample. Removal of human-derived reads was performed with Bowtie2 v2.3.5.1 on default settings by removing reads which aligned to the reference human genome, hg38.[51] This resulted in 257 samples with an average of 1,102,247 +/- 643,325 (σ) high quality, non-human reads per sample. Lastly, we used IGGSearch v1.0 on the 'lenient' preset (--min-reads-gene=1 –min-perc-genes=15 –min-sp-quality=25) to assign operational taxonomic units (OTU) to our quality-filtered sequences.[52] This produced 238 samples with an average of 24,888 +/- 16,340 (σ) high-quality, marker gene reads per sample.

Taxonomic Analysis:

Data analysis was performed using R v3.6.3. For all sequencing runs, a synthetic microbial community DNA standard (ZymoBIOMICS D6305) was included as a control. When necessary, the first step in our compositional analysis was filtering taxa, from all

samples, that contaminated the community standard control. Next, unassigned and mitochondrial reads were removed from our samples. Afterwards, we excluded 16S and ITS samples with fewer than 2,500 and 1,000 reads, respectively, as these samples did not have sufficient read depth to fully represent their microbial diversity (Supplementary Figure 1.11). Filtering was not required, nor performed for shotgun samples. The final number of 16S, ITS, and shotgun samples with high-quality, taxonomically assigned reads, was 147, 98, and 238, respectively (Supplementary Tables 2-4).

The alpha diversities for both amplicon and shotgun data were obtained using the 'diversity' and 'specnumber' functions from the Vegan v2.5-6 package, using the default parameters. Linear-mixed effect models (LME) were used for significance testing among alpha diversities to account for random effects, such as plate batching effects, and multiple measurements per individual using the nlme package, v3.1-148. For all datasets, beta diversities were obtained using the 'adonis' function in Vegan to generate Bray-Curtis distance matrices and perform PERMANOVA significance testing from compositional data. Beta diversity was visualized using non-metric multidimensional scaling (NMDS) ordination obtained from the 'metaMDS' function in Vegan. Matrix correlation was assessed using the 'mantel' function in Vegan.

Differential Abundance Testing:

Our primary focus with the first sample set was to compare the microbial compositions of different sample types within the same individual. Therefore, we used ANCOM v2.1 in R to test for differentially abundant microbes since it can account for multiple variables and random effects.[53] We used ANCOM with 'sample type' as our variable of interest (mucosal brushes vs. mucosal aspirates vs. colonoscopy lavage

aspirates) and the individual of origin as a random effect. Other parameters included 'p_adjust = FDR' to control for the false discovery rate, and significance was determined at < 0.05.

For shotgun data, our primary focus was to compare the microbial composition of different subject types (Polyp-free vs. TA-bearing vs. SP-bearing). We used a univariate Kruskal-Wallis (KW) test with independent hypotheses weighting (IHW). IHW increases power while controlling the false discovery weight by utilizing covariate data that are independent of the null hypothesis.[54] Before testing, we excluded samples with 'Unknown/Other' subject types, and filtered taxa that were not present in at least one-third of samples. We also eliminated repeated measurements by averaging the microbial relative abundances of left and right mucosal aspirates from the same individual. Kruskal-Wallis tests were performed for each taxon with the subject type as the variable. The IHW v1.14.0 package was used to correct p-values for the false discovery rate, using the sum of read counts per taxon across all samples as our covariate. FDR-adjusted p-values < 0.05 were considered significant. When visualizing relative abundances using a $\log_{10}$ scale, a pseudo-count of 0.0001 was added to prevent the removal of samples containing zeroes.

Random Forests:

Random Forests (RF) were performed on shotgun-sequenced mucosal aspirates to determine if the subject type of a sample could be predicted based on microbial composition. To do this, we used the rfPermute v1.9.3 package in R. We began by filtering taxa which were not present in at least one-third of mucosal aspirate samples. Two-thirds of the 156 shotgun mucosal aspirates were used for training the RF classifiers, while the remaining one-third was used for testing our RF models. RfPermute parameters were set to

importance = TRUE, nrep = 100, ntree = 501, and mtry = 8. Afterwards, we generated receiver-operator curves (ROC) using the 'roc' function with default settings (pROC v1.18.0 package). Variables of importance were visualized with the 'VarImpPlot' function in the rfPermute package.

Pathway Enrichment Analysis:

Pathway enrichment analysis was done using unassembled shotgun reads with HUMAnN v3.0.1.[55] The program was ran using the default parameters and the ChocoPhlAn v296 and UniRef90 v201901b databases were used for alignment. The 'humann_renorm_table' and 'humann_join_tables' functions were used to create a pathway abundance matrix of normalized counts in copies per million. Significantly enriched pathways between subject types were determined with a Kruskal-Wallis test using IHW. The false discovery rate was corrected for using the total sum of normalized counts per pathway as our covariate. Significance was determined at FDR < 0.05. Z-scores were calculated from pathway abundances, and then were visualized on a heatmap generated by the 'dist' and 'hclust' functions in R.

Functional Metagenomic Analysis:

Analysis of individual microbial genes was performed by cross-assembling reads into contiguous sequences using MEGAHIT v1.1.1.[56] Contigs smaller than 2,500 bp were discarded and the remainder had open reading frames (ORFs) identified by Prodigal v2.6.3.[57] The resulting ORFs were functionally annotated using eggNOG mapper v2.0, using the eggNOG v5.0 database. [58] Individual samples were aligned to annotated ORFs using Bowtie2 v2.3 to obtain per-sample ORF abundances. Per sample ORF abundances were compiled into a single ORF abundance table using the 'pileup.sh' script from BBMap. ORF

counts were normalized to reads per kilobase per genome equivalent using MicrobeCensus v1.1.1 on default settings.[59] Principal coordinate analysis was performed using the 'cmdscale' function from Vegan to visualize the functional metagenome composition among sample and subject types. PERMANOVA and differential abundance testing were performed in the same manner as with taxonomy.

DECLARATIONS

Data availability statement:

Sequencing data is available on the Sequence Read Archive under the BioProject ID, PRJNA745329. The source data used to generate Figures 1.2, 1.3b-d, 1.4, and 1.5 are provided as a Source Data File. The ASV/OTU tables and corresponding metadata used to generate the source data can be found in 'Supplementary_File_2.xlsx'. Larger files used in our functional metagenomic analysis are available on the Dryad Digital Repository (https://doi.org/10.7280/D1J10M). Additional data and materials are available upon reasonable request.

Code availability statement:

All code for data processing and analysis are available on GitHub at: https://github.com/Javelarb/ACS_polyp_study.

Acknowledgements:

We would like to thank Claudia Weihe and Jennifer B.H. Martiny for allowing us to borrow laboratory equipment and giving insightful feedback, Andrew Oliver and Jason A. Rothman for their bioinformatic expertise, Clark Hendrickson for his assistance with sample preparation, and Heather Maughan for her helpful edits and suggestions. This study was

funded by institutional research grant #IRG-16-187-13 from the American Cancer Society, and J.A.-B. was supported by the NIH-IMSD training grant number GM055246.

Disclosures of interest:

The authors declare no competing or conflicts of interest.

Author contributions:

We would like to thank Claudia Weihe and Jennifer B.H. Martiny for allowing us to borrow laboratory equipment and giving insightful feedback, Andrew Oliver and Jason A. Rothman for their bioinformatic expertise, Clark Hendrickson for his assistance with sample preparation, and Heather Maughan for her helpful edits and suggestions. This study was funded by institutional research grant #IRG-16-187-13 from the American Cancer Society, and J.A.-B. was supported by the NIH-IMSD training grant number GM055246.

Ethics approval:

This study was approved by the Institutional Review Board (IRB) of the University of California, Irvine (HS# 2017-3869).

REFERENCES

1. Sung, Hyuna *et al.* "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians* 71, no. 3: 209–49. https://doi.org/10.3322/caac.21660. (May 2021)
2. Stoffel, Elena M. *Et al*. "Hereditary Colorectal Cancer Syndromes: American Society of Clinical Oncology Clinical Practice Guideline Endorsement of the Familial Risk–Colorectal Cancer: European Society for Medical Oncology Clinical Practice Guidelines." *Journal of Clinical Oncology* 33, no. 2: 209–17. https://doi.org/10.1200/JCO.2014.58.1322. (January 10, 2015)
3. Collins, S.M., E. Denou, E.F. Verdu, and P. Bercik. "The Putative Role of the Intestinal Microbiota in the Irritable Bowel Syndrome." *Digestive and Liver Disease* 41, no. 12: 850–53. https://doi.org/10.1016/j.dld.2009.07.023. (December 2009)
4. Verdam, Froukje J. *Et al*. "Human Intestinal Microbiota Composition Is Associated with Local and Systemic Inflammation in Obesity: Obese Gut Microbiota and Inflammation." *Obesity* 21, no. 12: E607–15. https://doi.org/10.1002/oby.20466. (December 2013)
5. Song, Mingyang, Wendy S. Garrett, and Andrew T. Chan. "Nutrients, Foods, and Colorectal Cancer Prevention." *Gastroenterology* 148, no. 6: 1244-1260.e16. https://doi.org/10.1053/j.gastro.2014.12.035. (May 2015)
6. "Diet, Nutrition, Physical Activity, and Colorectal Cancer." World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report. dietandcancerreport.org. (2018)

7. David, Lawrence A. *Et al*. "Diet Rapidly and Reproducibly Alters the Human Gut Microbiome." *Nature* 505, no. 7484: 559–63. https://doi.org/10.1038/nature12820. (January 2014)
8. Engen, Phillip A., Stefan J. Green, Robin M. Voigt, Christopher B. Forsyth, and Ali Keshavarzian. "The Gastrointestinal Microbiome: Alcohol Effects on the Composition of Intestinal Microbiota." *Alcohol Research: Current Reviews* 37, no. 2: 223–36. (2015)
9. Ley, Ruth E. "Obesity and the Human Microbiome:" *Current Opinion in Gastroenterology* 26, no. 1: 5–11. https://doi.org/10.1097/MOG.0b013e328333d751. (January 2010)
10. Mailing, Lucy J., Jacob M. Allen, Thomas W. Buford, Christopher J. Fields, and Jeffrey A. Woods. "Exercise and the Gut Microbiome: A Review of the Evidence, Potential Mechanisms, and Implications for Human Health." *Exercise and Sport Sciences Reviews* 47, no. 2: 75–85. https://doi.org/10.1249/JES.0000000000000183. (April 2019)
11. Nakanishi, Yuki, Maria T. Diaz-Meco, and Jorge Moscat. "Serrated Colorectal Cancer: The Road Less Travelled?" *Trends in Cancer* 5, no. 11: 742–54. https://doi.org/10.1016/j.trecan.2019.09.004. (November 2019)
12. Pino, Maria S., and Daniel C. Chung. "The Chromosomal Instability Pathway in Colon Cancer." *Gastroenterology* 138, no. 6: 2059–72. https://doi.org/10.1053/j.gastro.2009.12.065. (May 2010)
13. De Palma, Fatima *et al*. "The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer." *Cancers* 11, no. 7: 1017. https://doi.org/10.3390/cancers11071017. (July 20, 2019)
14. DeDecker, Lauren, Bretton Coppedge, Julio Avelar-Barragan, William Karnes, and Katrine Whiteson. "Microbiome Distinctions between the CRC Carcinogenic Pathways."

Gut Microbes, 1–12. https://doi.org/10.1080/19490976.2020.1854641. (January 15, 2021)

15. Kahi, Charles J. "Screening Relevance of Sessile Serrated Polyps." *Clinical Endoscopy* 52, no. 3: 235–38. https://doi.org/10.5946/ce.2018.112. (May 31, 2019)
16. Delker, Don A. *Et al*. "RNA Sequencing of Sessile Serrated Colon Polyps Identifies Differentially Expressed Genes and Immunohistochemical Markers." Edited by Frank T. Kolligs. *PloS ONE* 9, no. 2: e88367. https://doi.org/10.1371/journal.pone.0088367. (February 12, 2014)
17. Peters, Brandilyn A. *Et al*. "The Gut Microbiota in Conventional and Serrated Precursors of Colorectal Cancer." *Microbiome* 4, no. 1: 69. https://doi.org/10.1186/s40168-016-0218-6. (December 2016)
18. Yoon, Hyuk *et al.* "Comparisons of Gut Microbiota Among Healthy Control, Patients With Conventional Adenoma, Sessile Serrated Adenoma, and Colorectal Cancer." *Journal of Cancer Prevention* 22, no. 2: 108–14. https://doi.org/10.15430/JCP.2017.22.2.108. (June 30, 2017)
19. Rezasoltani, Sama, *et al*. "The Association between Fecal Microbiota and Different Types of Colorectal Polyp as Precursors of Colorectal Cancer." *Microbial Pathogenesis* 124: 244–49. https://doi.org/10.1016/j.micpath.2018.08.035. (November 2018)
20. Riva, Alessandra *et al*. "A Fiber-Deprived Diet Disturbs the Fine-Scale Spatial Architecture of the Murine Colon Microbiome." *Nature Communications* 10, no. 1: 4366. https://doi.org/10.1038/s41467-019-12413-0. (December 2019)
21. Chen, Weiguang, Fanlong Liu, Zongxin Ling, Xiaojuan Tong, and Charlie Xiang. "Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal

Cancer." Edited by Antonio Moschetta. *PloS ONE* 7, no. 6: e39743. https://doi.org/10.1371/journal.pone.0039743. (June 28, 2012)

22. Wu, G. D. *Et al*. "Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes." *Science* 334, no. 6052: 105–8. https://doi.org/10.1126/science.1208344. (October 7, 2011)
23. Turnbaugh, P. J. *Et al*. "The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice." *Science Translational Medicine* 1, no. 6: 6ra14-6ra14. https://doi.org/10.1126/scitranslmed.3000322. (November 11, 2009)
24. Hall, Andrew Brantley *et al*. "A Novel Ruminococcus Gnavus Clade Enriched in Inflammatory Bowel Disease Patients." *Genome Medicine* 9, no. 1: 103. https://doi.org/10.1186/s13073-017-0490-5. (November 28, 2017)
25. Ridlon, Jason M., and Phillip B. Hylemon. "Identification and Characterization of Two Bile Acid Coenzyme A Transferases from Clostridium Scindens, a Bile Acid 7α-Dehydroxylating Intestinal Bacterium." *Journal of Lipid Research* 53, no. 1: 66–76. https://doi.org/10.1194/jlr.M020313. (January 2012)
26. Marion, Solenne *et al*. "In Vitro and in Vivo Characterization of Clostridium Scindens Bile Acid Transformations." *Gut Microbes* 10, no. 4: 481–503. https://doi.org/10.1080/19490976.2018.1549420. (July 4, 2019)
27. Ajouz, Hana, Deborah Mukherji, and Ali Shamseddine. "Secondary Bile Acids: An Underrecognized Cause of Colon Cancer." *World Journal of Surgical Oncology* 12, no. 1: 164. https://doi.org/10.1186/1477-7819-12-164. (2014)

28. Ocvirk, Soeren, and Stephen J.D. O'Keefe. "Dietary Fat, Bile Acid Metabolism and Colorectal Cancer." *Seminars in Cancer Biology*, S1044579X2030208X. https://doi.org/10.1016/j.semcancer.2020.10.003. (October 2020)
29. Haghi, Fakhri, Elshan Goli, Bahman Mirzaei, and Habib Zeighami. "The Association between Fecal Enterotoxigenic B. Fragilis with Colorectal Cancer." *BMC Cancer* 19, no. 1: 879. https://doi.org/10.1186/s12885-019-6115-1. (December 2019)
30. Ulger Toprak, N. *Et al*. "A Possible Role of Bacteroides Fragilis Enterotoxin in the Aetiology of Colorectal Cancer." *Clinical Microbiology and Infection* 12, no. 8: 782–86. https://doi.org/10.1111/j.1469-0691.2006.01494.x. (August 2006)
31. Cheng, Wai Teng, Haresh Kumar Kantilal, and Fabian Davamani. "The Mechanism of Bacteroides Fragilis Toxin Contributes to Colon Cancer Formation." *The Malaysian Journal of Medical Sciences: MJMS* 27, no. 4: 9–21. https://doi.org/10.21315/mjms2020.27.4.2. (July 2020)
32. Gholizadeh, Pourya, Hosein Eslami, and Hossein Samadi Kafil. "Carcinogenesis Mechanisms of Fusobacterium Nucleatum." *Biomedicine & Pharmacotherapy* 89: 918–25. https://doi.org/10.1016/j.biopha.2017.02.102. (May 2017)
33. Bess, Elizabeth N. *Et al*. "Genetic Basis for the Cooperative Bioactivation of Plant Lignans by Eggerthella Lenta and Other Human Gut Bacteria." *Nature Microbiology* 5, no. 1: 56–66. https://doi.org/10.1038/s41564-019-0596-1. (January 2020)
34. Webb, Amy L., and Marjorie L. McCullough. "Dietary Lignans: Potential Role in Cancer Prevention." *Nutrition and Cancer* 51, no. 2: 117–31. https://doi.org/10.1207/s15327914nc5102_1. (March 2005)

35. Adlercreutz, Herman. “Lignans and Human Health.” *Critical Reviews in Clinical Laboratory Sciences* 44, no. 5–6: 483–525. https://doi.org/10.1080/10408360701612942. (January 2007)

36. Aune, D. *Et al*. “Dietary Fibre, Whole Grains, and Risk of Colorectal Cancer: Systematic Review and Dose-Response Meta-Analysis of Prospective Studies.” *BMJ* 343, no. Nov10 1: d6617–d6617. https://doi.org/10.1136/bmj.d6617. (November 11, 2011)

37. Donohoe, Dallas R. *Et al*. “The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon.” *Cell Metabolism* 13, no. 5: 517–26. https://doi.org/10.1016/j.cmet.2011.02.018. (May 4, 2011)

38. Hague, Angela, Douglas J. E. Elder, Diane J. Hicks, and Christos Paraskeva. “Apoptosis in Colorectal Tumour Cells: Induction by the Short Chain Fatty Acids Butyrate, Propionate and Acetate and by the Bile Salt Deoxycholate.” *International Journal of Cancer* 60, no. 3: 400–406. https://doi.org/10.1002/ijc.2910600322. (January 27, 1995)

39. Hamer, H. M. *Et al*. “Review Article: The Role of Butyrate on Colonic Function.” *Alimentary Pharmacology & Therapeutics* 27, no. 2: 104–19. https://doi.org/10.1111/j.1365-2036.2007.03562.x. (October 26, 2007)

40. Davie, James R. “Inhibition of Histone Deacetylase Activity by Butyrate.” *The Journal of Nutrition* 133, no. 7: 2485S-2493S. https://doi.org/10.1093/jn/133.7.2485S. (July 1, 2003)

41. Goldstein, Neal S. “Serrated Pathway and APC (Conventional)-Type Colorectal Polyps: Molecular-Morphologic Correlations, Genetic Pathways, and Implications for Classification.” *American Journal of Clinical Pathology* 125, no. 1: 146–53. https://doi.org/10.1309/87BD0C6UCGUG236J. (January 2006)

42. Allen-Vercoe et al., "Anaerostipes Hadrus Comb. Nov., a Dominant Species within the Human Colonic Microbiota; Reclassification of Eubacterium Hadrum Moore et al. 1976." *Anaerobe* 18, no. 5: 523-529. 10.1016/j.anaerobe.2012.09.002. (October 2012)
43. Looby, Caitlin I., Mia R. Maltz, and Kathleen K. Treseder. "Belowground Responses to Elevation in a Changing Cloud Forest." *Ecology and Evolution* 6, no. 7: 1996–2009. https://doi.org/10.1002/ece3.2025. (April 2016)
44. Weihe, Claudia, and Avelar-Barragan, Julio. "Next Generation Shotgun Library Preparation for Illumina Sequencing – Low Volume V1." https://doi.org/10.17504/protocols.io.bvv8n69w (2021)
45. Bolyen, Evan *et al*. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37, no. 8: 852–57. https://doi.org/10.1038/s41587-019-0209-9. (August 2019)
46. Callahan, Benjamin J *et al*. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13, no. 7: 581–83. https://doi.org/10.1038/nmeth.3869. (July 2016)
47. McDonald, Daniel *et al*. "An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea." *The ISME Journal* 6, no. 3: 610–18. https://doi.org/10.1038/ismej.2011.139. (March 2012)
48. Nilsson, Rolf Henrik *et al*. "The UNITE Database for Molecular Identification of Fungi: Handling Dark Taxa and Parallel Taxonomic Classifications." *Nucleic Acids Research* 47, no. D1: D259–64. https://doi.org/10.1093/nar/gky1022. (January 8, 2019)
49. Bushnell, Brian. "BBMap: A Fast, Accurate, Splice-Aware Aligner." Lawrence Berkeley National Lab. (2014)

50. Cantu, Vito Adrian, Jeffrey Sadural, and Robert Edwards. "PRINSEQ++, a Multi-Threaded Tool for Fast and Efficient Quality Control and Preprocessing of Sequencing Datasets." Preprint. PeerJ Preprints, https://doi.org/10.7287/peerj.preprints.27553v1. (February 27, 2019)
51. Langmead, Ben, and Steven L Salzberg. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9, no. 4: 357–59. https://doi.org/10.1038/nmeth.1923. (April 2012)
52. Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." *Nature* 568, no. 7753: 505–10. https://doi.org/10.1038/s41586-019-1058-x. (April 2019)
53. Mandal, Siddhartha *et al*. "Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition." *Microbial Ecology in Health & Disease* 26, no. 0. https://doi.org/10.3402/mehd.v26.27663. (May 29, 2015)
54. Ignatiadis, Nikolaos, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. "Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing." *Nature Methods* 13, no. 7: 577–80. https://doi.org/10.1038/nmeth.3885. (July 2016)
55. Beghini, Francesco *et al*. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with BioBakery 3." *Elife* 10: e65088. https://doi.org/10.7554/eLife.65088. (May 4, 2021)
56. Li, Dinghua *et al*. "MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven by Advanced Methodologies and Community Practices." *Methods* 102: 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020. (June 2016)

57. Hyatt, Doug *et al*. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11, no. 1: 119. https://doi.org/10.1186/1471-2105-11-119. (December 2010)

58. Huerta-Cepas, Jaime *et al*. "EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47, no. D1: D309–14. https://doi.org/10.1093/nar/gky1085. (January 8, 2019)

59. Nayfach, Stephen, and Katherine S Pollard. "Average Genome Size Estimation Improves Comparative Metagenomics and Sheds Light on the Functional Ecology of the Human Microbiome." *Genome Biology* 16, no. 1: 51. https://doi.org/10.1186/s13059-015-0611-7. (December 2015)

TABLES & FIGURES

Table 1.1*: Study cohort information.* A table describing the sample sizes, sample types, median age, median BMI, ethnicity compositions, and sex ratios of each sample set. The first sample set was sequenced twice, once using 16S sequencing and once using ITS sequencing.

	Sample set 1 (16S)	**Sample set 1 (ITS)**	**Sample set 2 (Shotgun)**
Number of samples	147	98	238
Sample types	Mucosal brushes Mucosal aspirates Lavage aspirates	Mucosal brushes Mucosal aspirates Lavage aspirates	Mucosal aspirates Lavage aspirates Fecal samples
Median Age (Years)	60	61	65
Median BMI (kg/m^2)	25	25	26
Ethnicity	White: 60% Black: 7% Asian: 21% Hispanic: 8% Other/Unknown: 4%	White: 71% Black: 3% Asian: 13% Hispanic: 11% Other/Unknown: 2%	White: 58% Black: 1% Asian: 16% Hispanic: 11% Other/Unknown: 14%
Sex	Male: 57% Female: 43% Other/Unknown: 0%	Male: 63% Female: 37% Other/Unknown: 0%	Male: 48% Female: 39% Other/Unknown: 13%

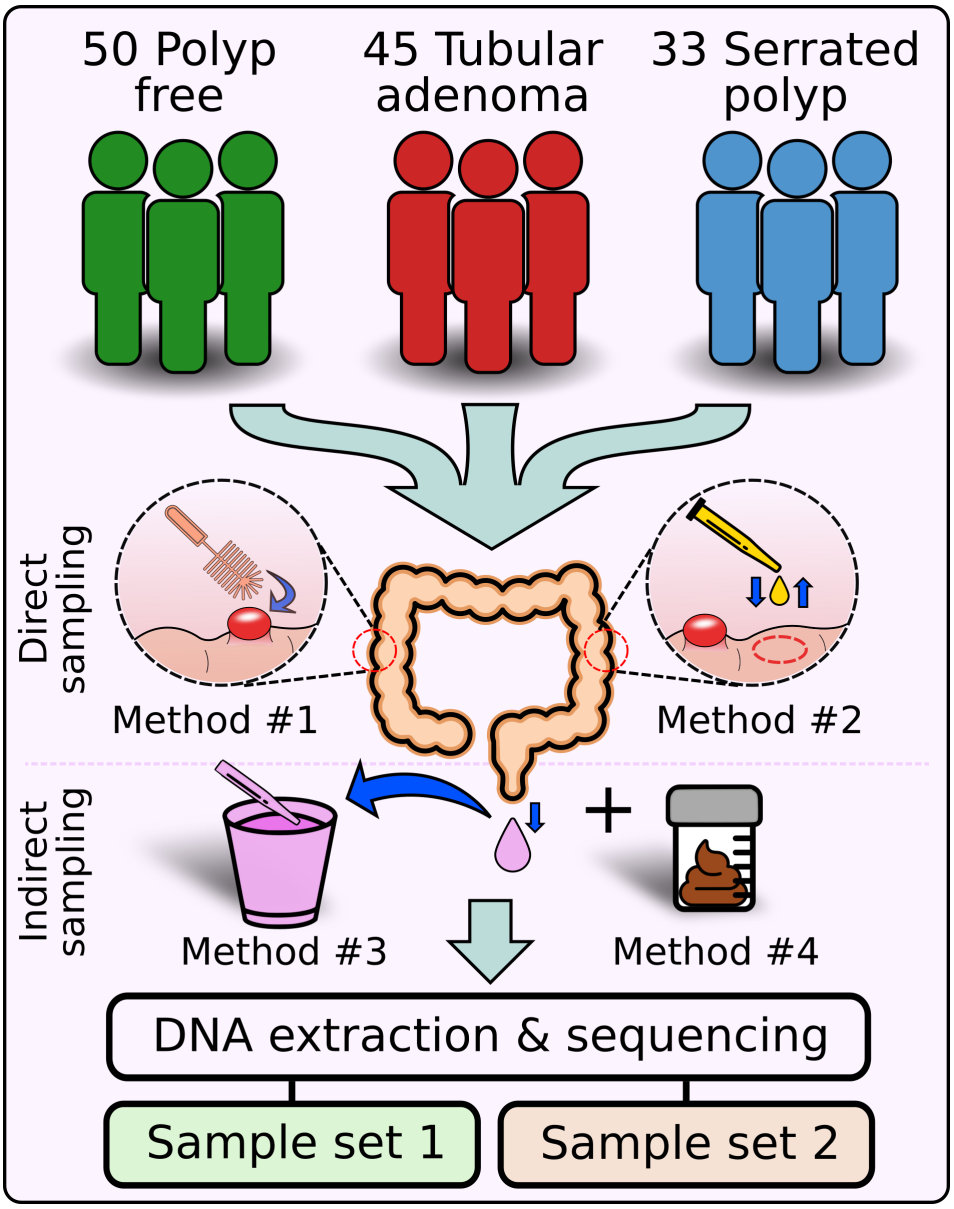


Figure 1.1: Study design.

A total of 140 individuals were recruited for this study, including 50 polyp-free individuals, 45 with tubular adenomas, and 33 with serrated polyps (HPP, TSA, or SSP). The remaining 12 individuals had missing or unknown pathologies. Multiple samples were taken from each subject during colonoscopy. This included mucosal brushes (Method #1, orange), mucosal aspirates (Method #2, yellow), and lavage aspirates (Method #3, purple). Fecal samples (Method #4, brown) were collected from participants four to six weeks post-

colonoscopy. DNA extraction and sequencing produced two sample sets. The first sample set was produced by sequencing mucosal brushes, mucosal aspirates, and lavage aspirates using 16S and ITS sequencing. The second sample set was produced by sequencing mucosal aspirates, lavage aspirates, and fecal samples using whole-genome shotgun sequencing.

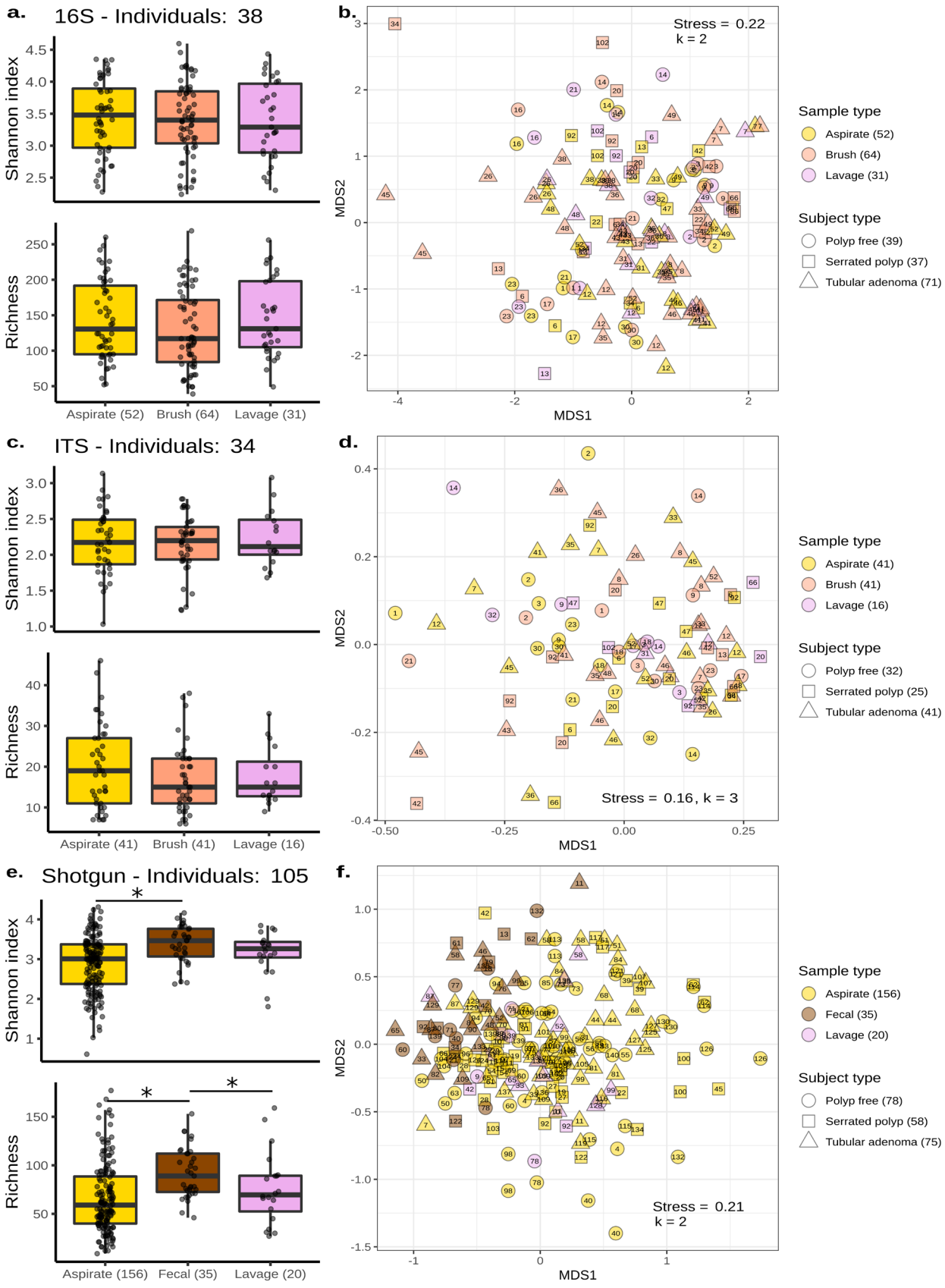

a.
16S - Individuals: 38
Shannon index
Richness
Aspirate (52)
Brush (64)
Lavage (31)
b.
Stress = 0.22
k = 2
MDS1
MDS2
Sample type
Aspirate (52)
Brush (64)
Lavage (31)
Subject type
Polyp free (39)
Serrated polyp (37)
Tubular adenoma (71)
c.
ITS - Individuals: 34
Shannon index
Richness
Aspirate (41)
Brush (41)
Lavage (16)
d.
Stress = 0.16, k = 3
MDS1
MDS2
Sample type
Aspirate (41)
Brush (41)
Lavage (16)
Subject type
Polyp free (32)
Serrated polyp (25)
Tubular adenoma (41)
e.
Shotgun - Individuals: 105
Shannon index
Richness
Aspirate (156)
Fecal (35)
Lavage (20)
f.
Stress = 0.21
k = 2
MDS1
MDS2
Sample type
Aspirate (156)
Fecal (35)
Lavage (20)
Subject type
Polyp free (78)
Serrated polyp (58)
Tubular adenoma (75)

Figure 1.2: Microbiomes of Mucosal and Lavage Samples are similar to each other but different from those in Feces.

A, C, and E) Box plots showing Shannon diversity and richness estimates across mucosal aspirates (yellow), mucosal brushes (orange), lavage aspirates (purple), and fecal samples (brown). The first sample set was sequenced using 16S **(A)**, and ITS **(C)** sequencing. The second sample set was sequenced using shotgun sequencing **(E)**. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. **B, D, and F)** Non-metric multidimensional scaling of Bray-Curtis dissimilarities produced from 16S **(B)**, ITS **(D)**, and shotgun **(F)** compositional data. Each point corresponds to one sample, with multiple samples per individual. The individual of origin is denoted numerically within each point. The number of samples per sample type and subject category are annotated parenthetically. Significant comparisons (Linear mixed effects: $p < 0.05$) are denoted with an asterisk (*). Source data are provided as a Source Data file.

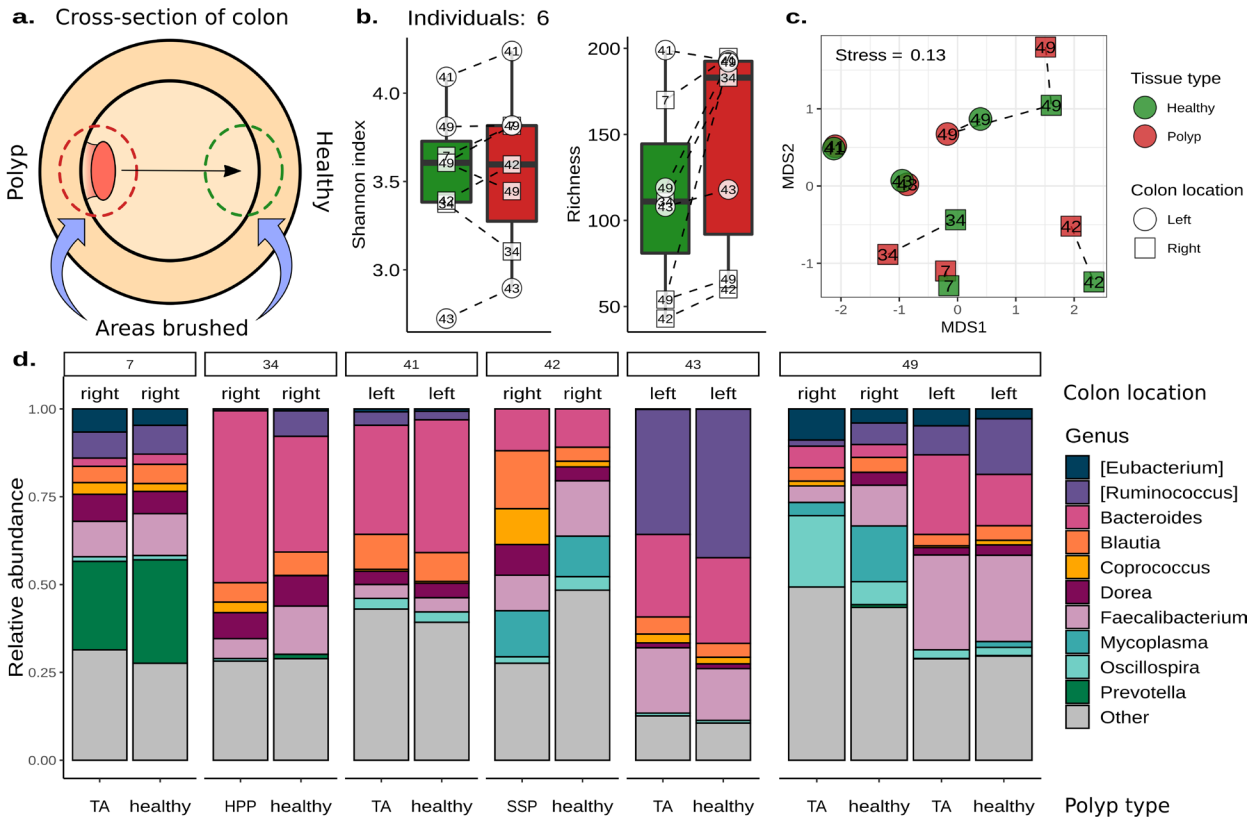


Figure 1.3: The Microbiomes of Polyps and Healthy Opposite Wall Tissue are similar within Individuals.

A) An illustration of the sampling strategy used to characterize the microbial community of 16S mucosal brushes from polyps (red) and healthy opposite wall tissue (green). **B)** Box plots of Shannon diversity and richness estimates from polyp and healthy opposite wall brushes. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. **C)** Non-metric multidimensional scaling of Bray-Curtis dissimilarities of polyp and healthy opposite wall tissue brushes. Each point is one sample, with multiple samples per individual. The individual of origin is denoted numerically within each point. The shape of each point denotes the right (proximal) and left (distal) side of the colon. **D)** The relative abundance of

the top ten microbial genera across all samples. Samples are grouped by each individual and labeled by polyp type, where tubular adenoma = TA, hyperplastic polyp = HPP, and sessile serrated polyp = SSP. Source data for Figures 3b-d are provided as a Source Data file.

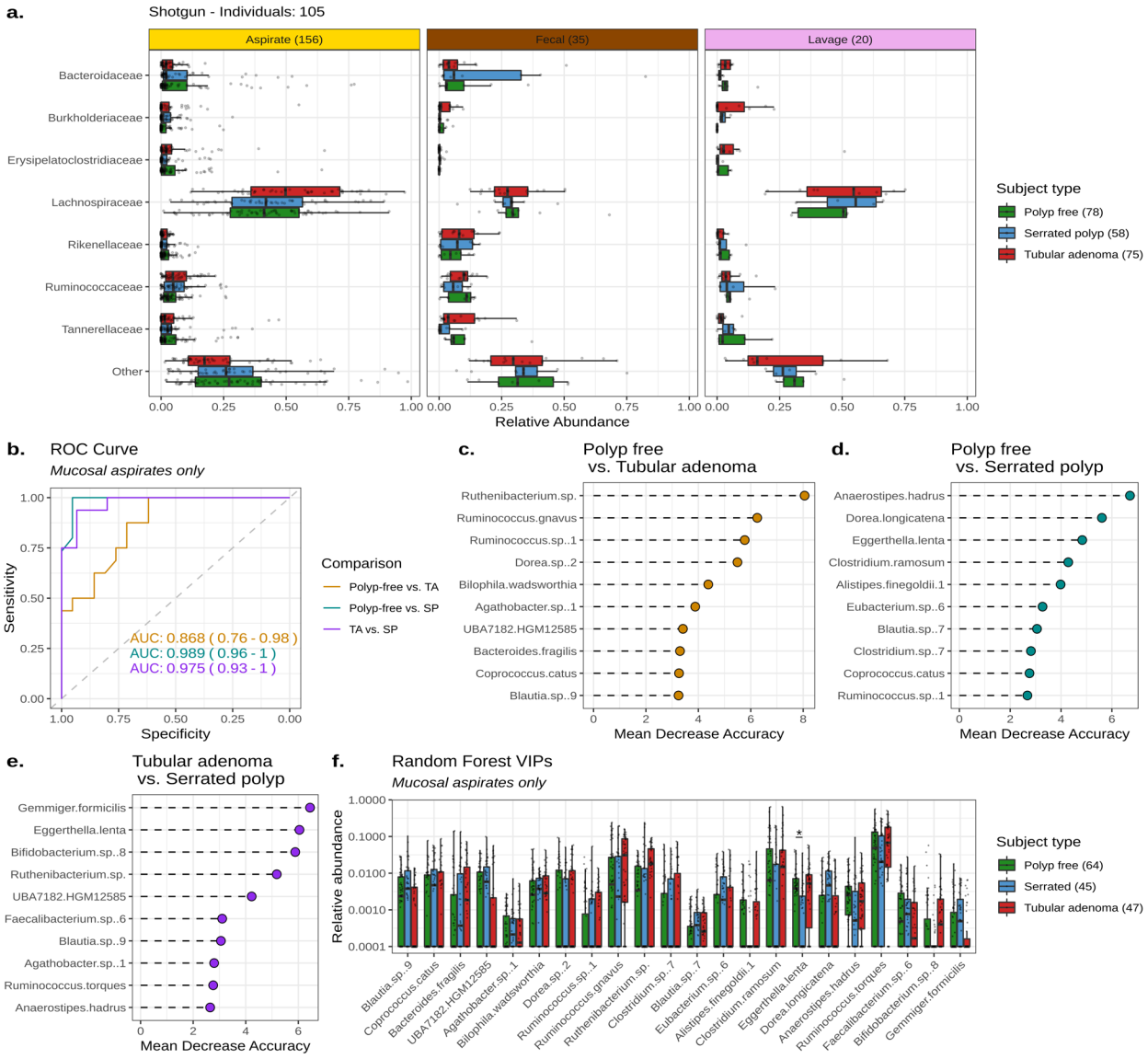


Figure 1.4: Tubular Adenoma-bearing, Serrated Polyp-bearing, and Polyp-free Individuals have distinct Microbiomes.

A) Box plots of the top seven most abundant microbial families across all samples from the second sample set. The number of samples per sampling method and subject type are denoted parenthetically, with multiple samples per individual. **B)** A receiver operating characteristic (ROC) curve illustrating the true positive rate (Sensitivity, y-axis) versus the false positive rate (Specificity, x-axis) produced by Random Forest classification of second

sample set mucosal aspirates. The area under the curve (AUC) value for each Random Forest is displayed with the 90% confidence interval. **C, D, and E)** The top ten variables of importance for each pairwise Random Forest classification. Variables are sorted by their mean decrease in accuracy, with larger means contributing greater to Random Forest performance. **F)** Box plots displaying the relative abundances of the top Random Forest variables of importance. Each point is one sample, with multiple samples per individual. A pseudo-count of 0.0001 was added to visualize samples which had relative abundances of zero, since the y axis is scaled to $\log_{10}$. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. Significant comparisons (Kruskal-Wallis: p-adj < 0.05) are denoted with an asterisk (*). Source data are provided as a Source Data file.

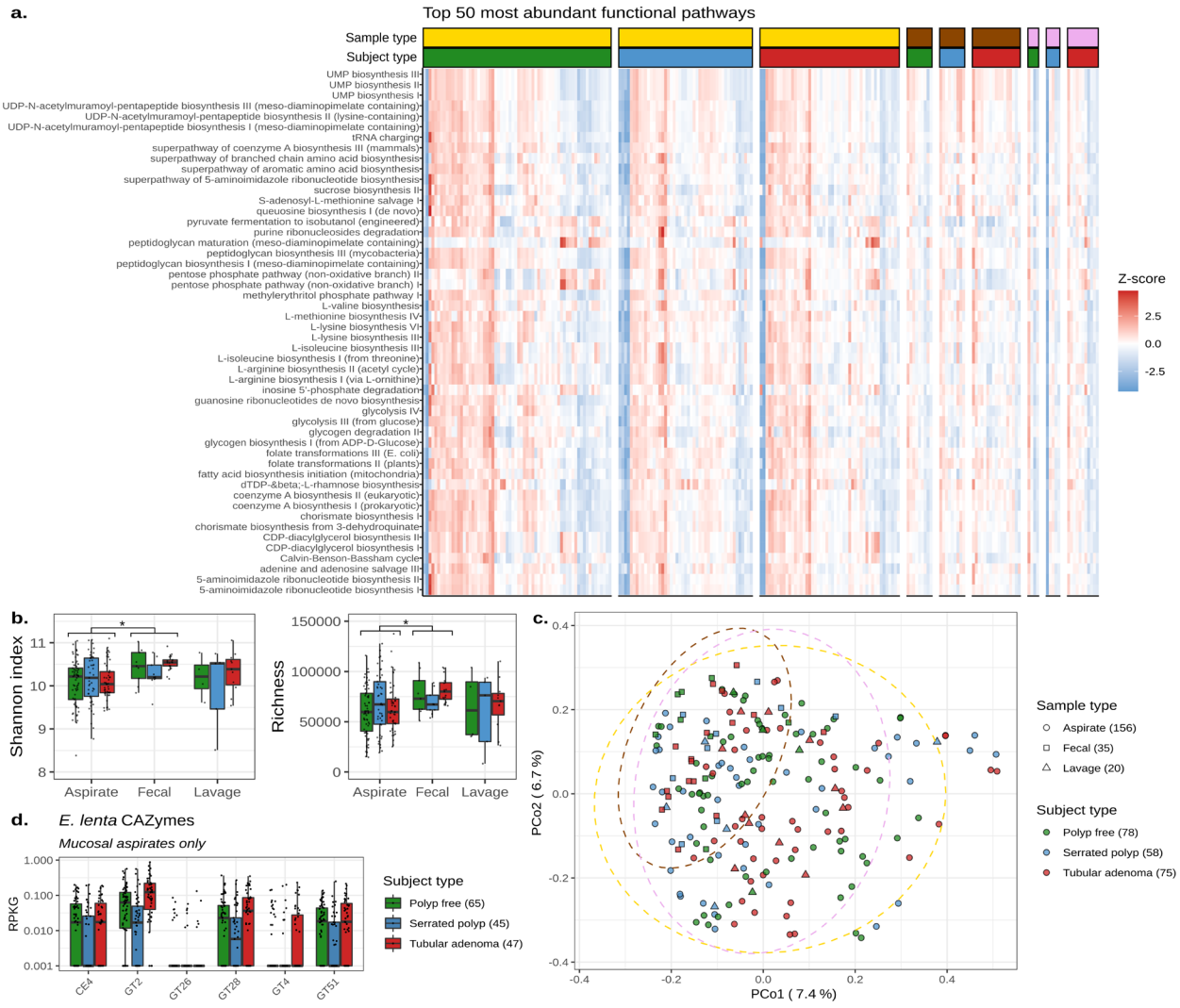
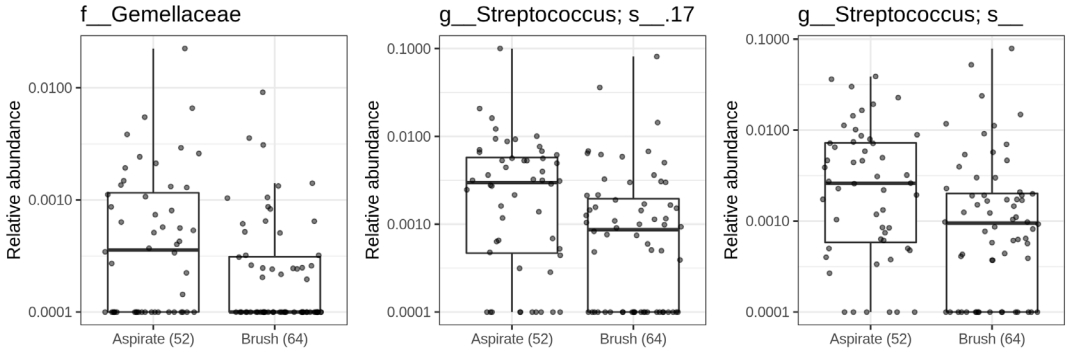


Figure 1.5: Microbiome Functional Potential is distinct across Sampling Methods and Subject Types.

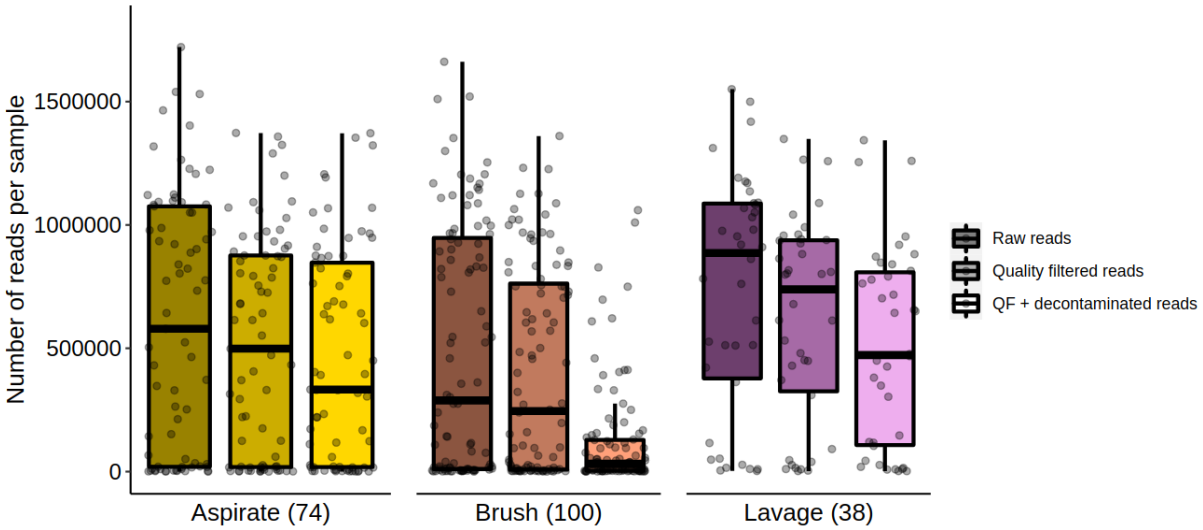
A) A heatmap displaying the z-scores of the top 50 most abundant microbial pathways found within the second sample set. Each column is one sample, with multiple samples per individual. Samples are clustered by sample and subject types. Yellow represents mucosal aspirates, brown represents fecal samples, and purple represents lavage aspirates. Within subject types, green represents polyp free samples, blue represents serrated polyp samples, and red represents tubular adenoma samples. **B)** Box plots showing the Shannon

diversity and richness of individual microbial genes across second sample set mucosal aspirates, lavage aspirates, and fecal samples. Significant comparisons (Linear mixed effects: $p < 0.05$) are denoted with an asterisk (*). **C)** Principal coordinate analysis of per-gene Bray-Curtis dissimilarities. Each point represents one sample. Ellipses are drawn to represent the 95% confidence interval of each sample type's distribution. The number of samples per sampling method and subject type are annotated parenthetically. **D)** Box plots showing the abundance of *E. lenta* specific carbohydrate active enzymes in reads per kilobase per genome equivalent. Only mucosal aspirates from the second sample set are shown, with the number of mucosal aspirates per subject type being denoted parenthetically. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. Source data are provided as a Source Data file.

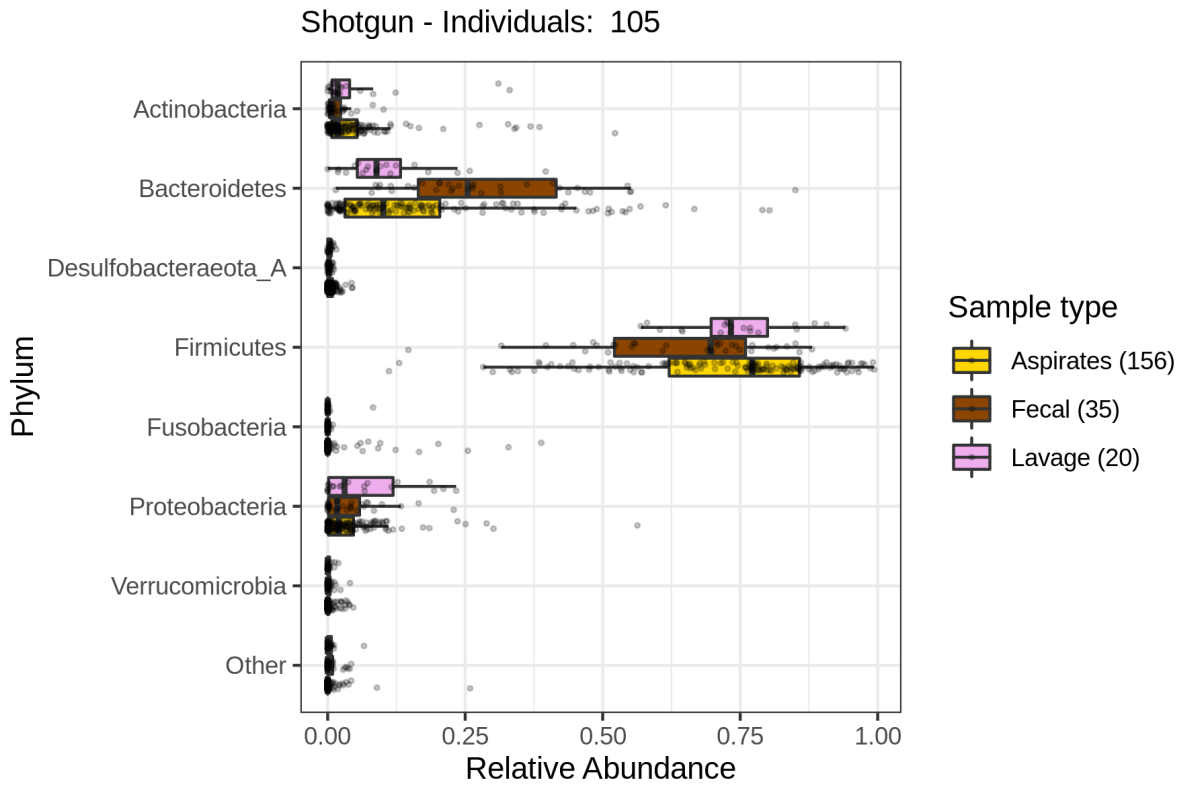
Supplementary Figures and Tables:



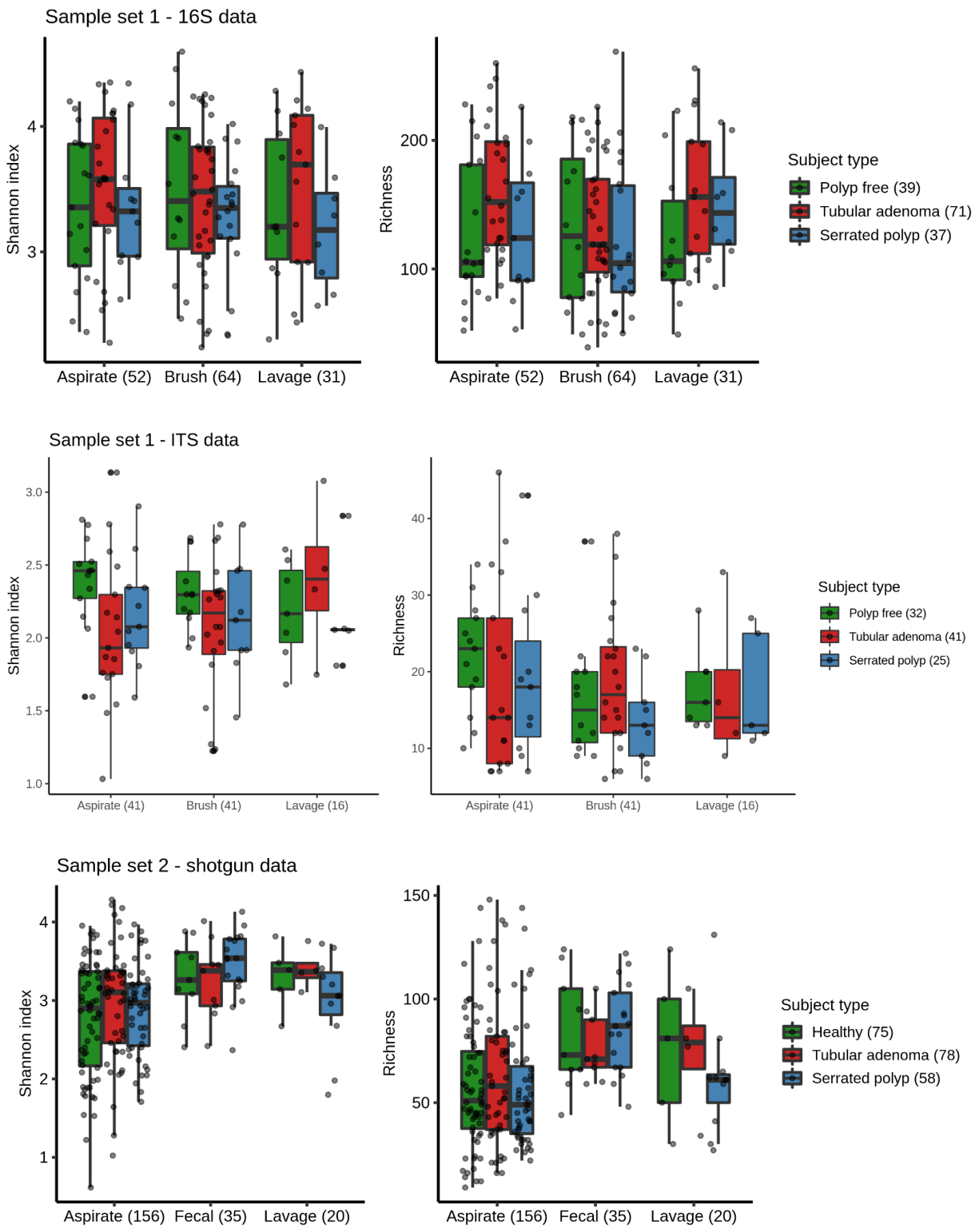
Supplementary Figure 1.1: Box plots displaying the relative abundance of microbes determined to be differentially abundant by ANCOM2 (p-adj < 0.05). Data is from mucosal brushes and mucosal aspirates from the first sample set. Each point is one sample, with multiple samples per individual. Plots are labeled with the most specific taxonomic rank for each ASV. A pseudo-count of 0.0001 was added to visualize samples which had relative abundances of zero, since the y axis is scaled to $\log_{10}$. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range.



Supplementary Figure 1.2: Box plots showing the number of reads per sample produced by a pilot shotgun sequencing run using mucosal brushes, mucosal aspirates, and lavage aspirates from the first sample set. Each point is one sample, with multiple samples per individual. The number of samples per sampling method is denoted parenthetically. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. 'Raw reads' refers to the number of reads produced by the Illumina NextSeq platform. 'Quality filtered reads' refers to the number of reads after removing reads with a quality score lower than a mean of 28. 'QF + decontaminated reads' refers to the number of reads after removing human-derived reads.

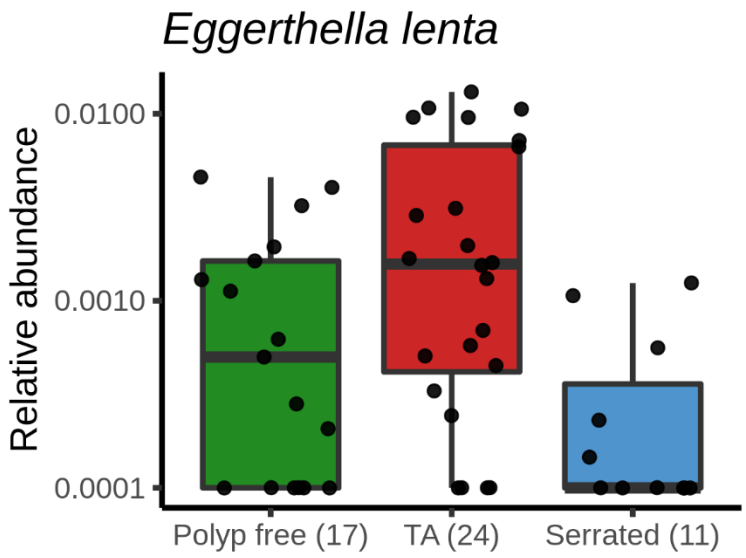


Supplementary Figure 1.3: Box plots showing the relative abundance of the top seven most abundant microbial phyla across mucosal aspirates, lavage aspirates, and fecal samples from the second sample set. Each point is one sample, with multiple samples per individual. The number of samples per sampling method is denoted parenthetically. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range.

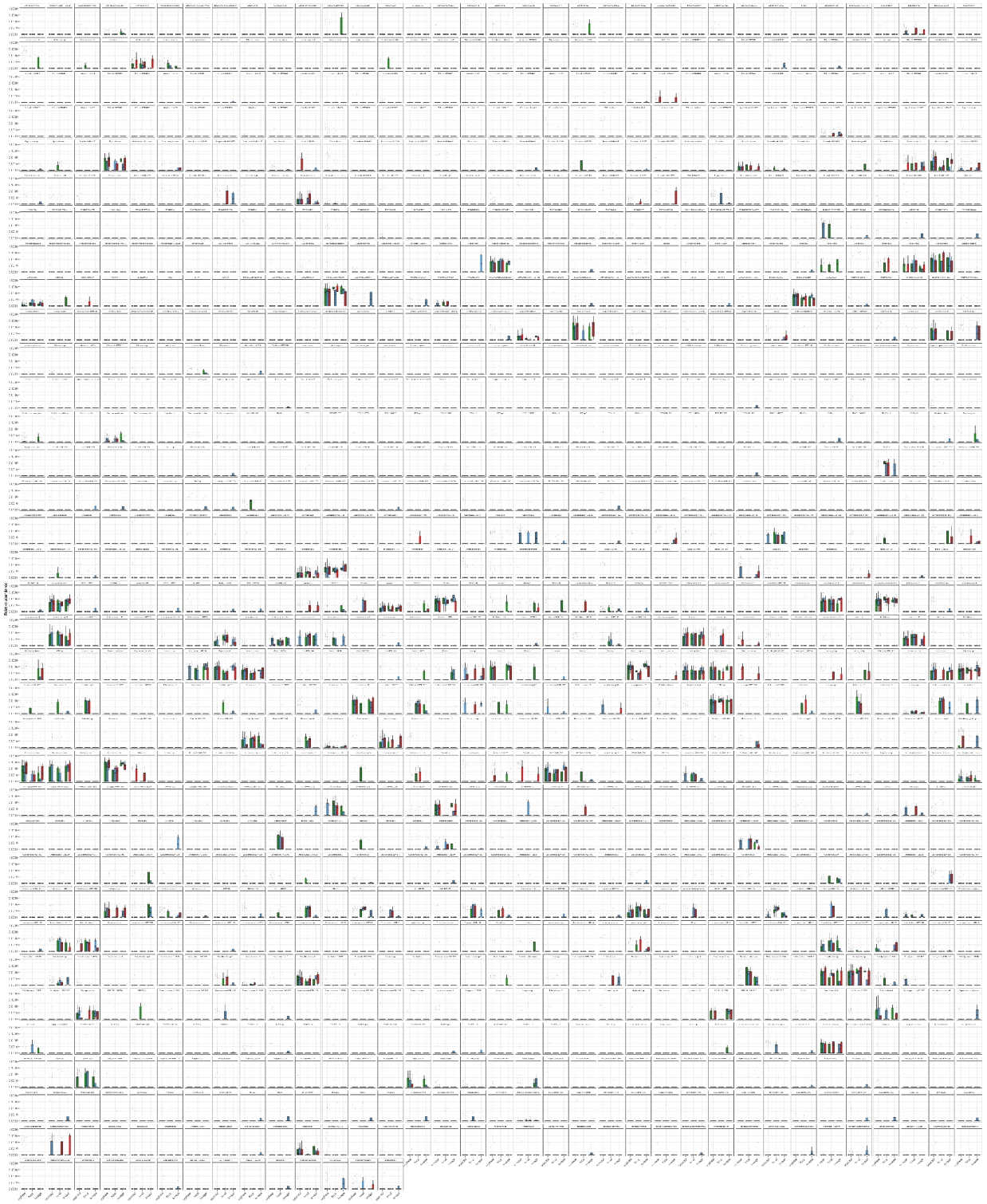


Supplementary Figure 1.4: Box plots showing Shannon diversity and richness estimates across the first and second sample sets. The number of samples for each sampling method

and subject type are denoted parenthetically. Each point is one sample, with multiple samples per individual. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. There was significantly increased Shannon diversity in polyp-free ITS samples when compared to TA samples (Linear mixed effects model: $p = 0.03$).

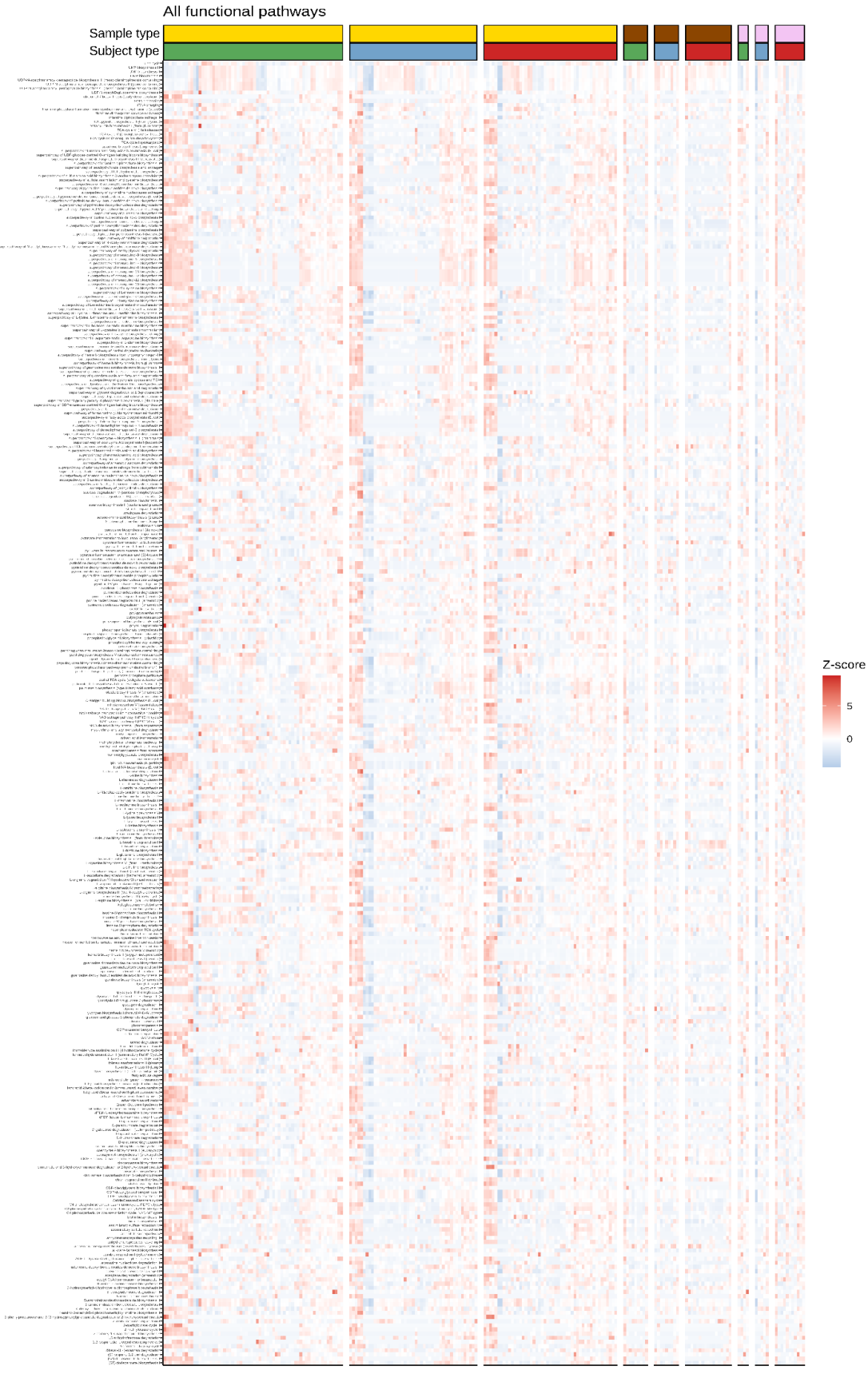


Supplementary Figure 1.5: A box plot showing the relative abundance of *E. lenta* in 16S mucosal aspirates from the first sample set across subject types. Each point is one sample, with multiple samples per individual. A pseudo-count of 0.0001 was added to visualize samples which had a relative abundance of zero, since the y-axis is scaled to $\log_{10}$. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range.

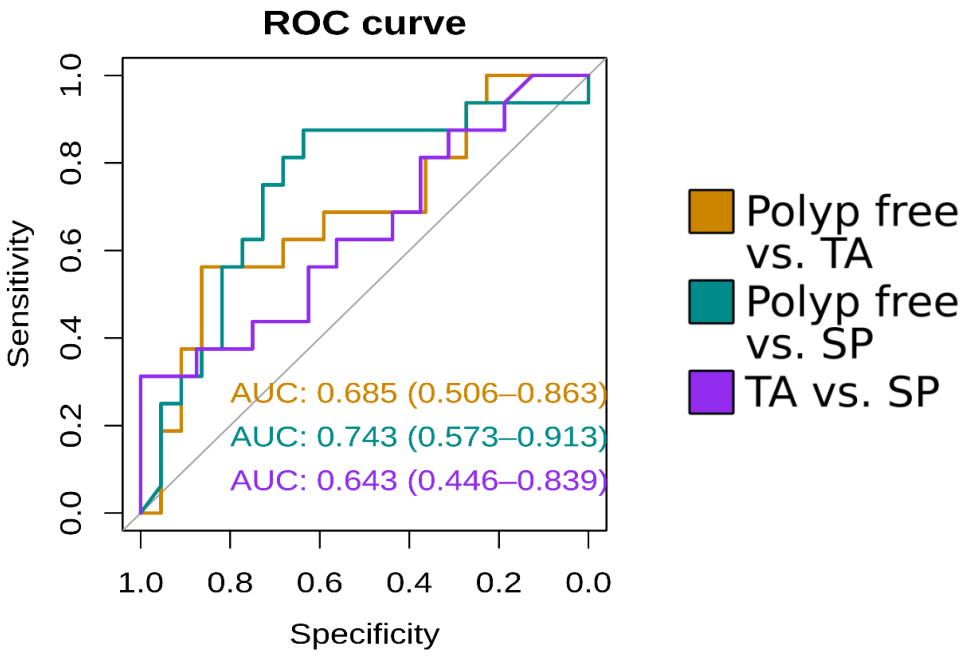


Supplementary Figure 1.6: Box plots showing the relative abundance of all OTUs from the second sample set. Each point is one sample, with multiple samples per individual. Samples

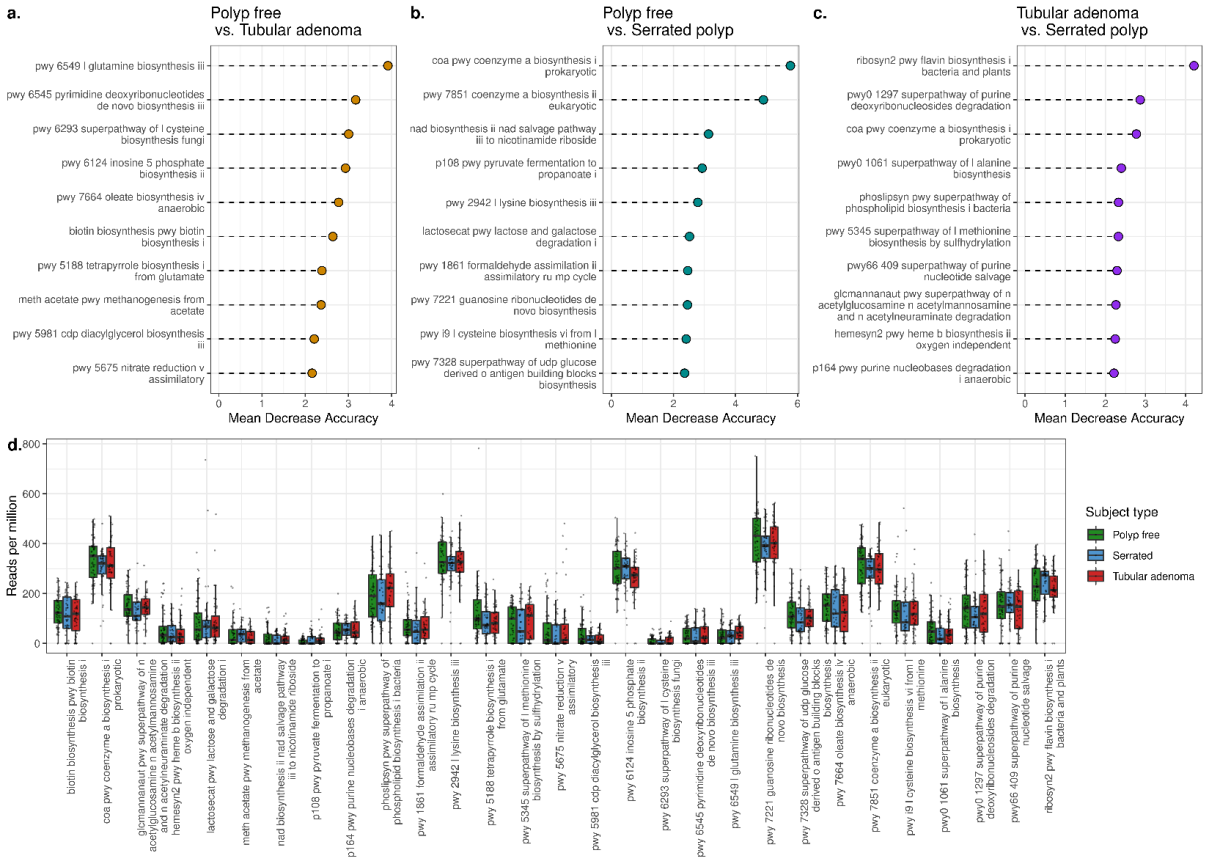
are faceted by sample type and colored by subject type. Green refers to polyp-free samples, red refers to TA-bearing samples, and blue refers to SP-bearing samples. A pseudo-count of 0.0001 was added to visualize samples which had a relative abundance of zero, since the y-axis is scaled to $\log_{10}$. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range.


All functional pathways
Sample type
Subject type
Z-score
5
0

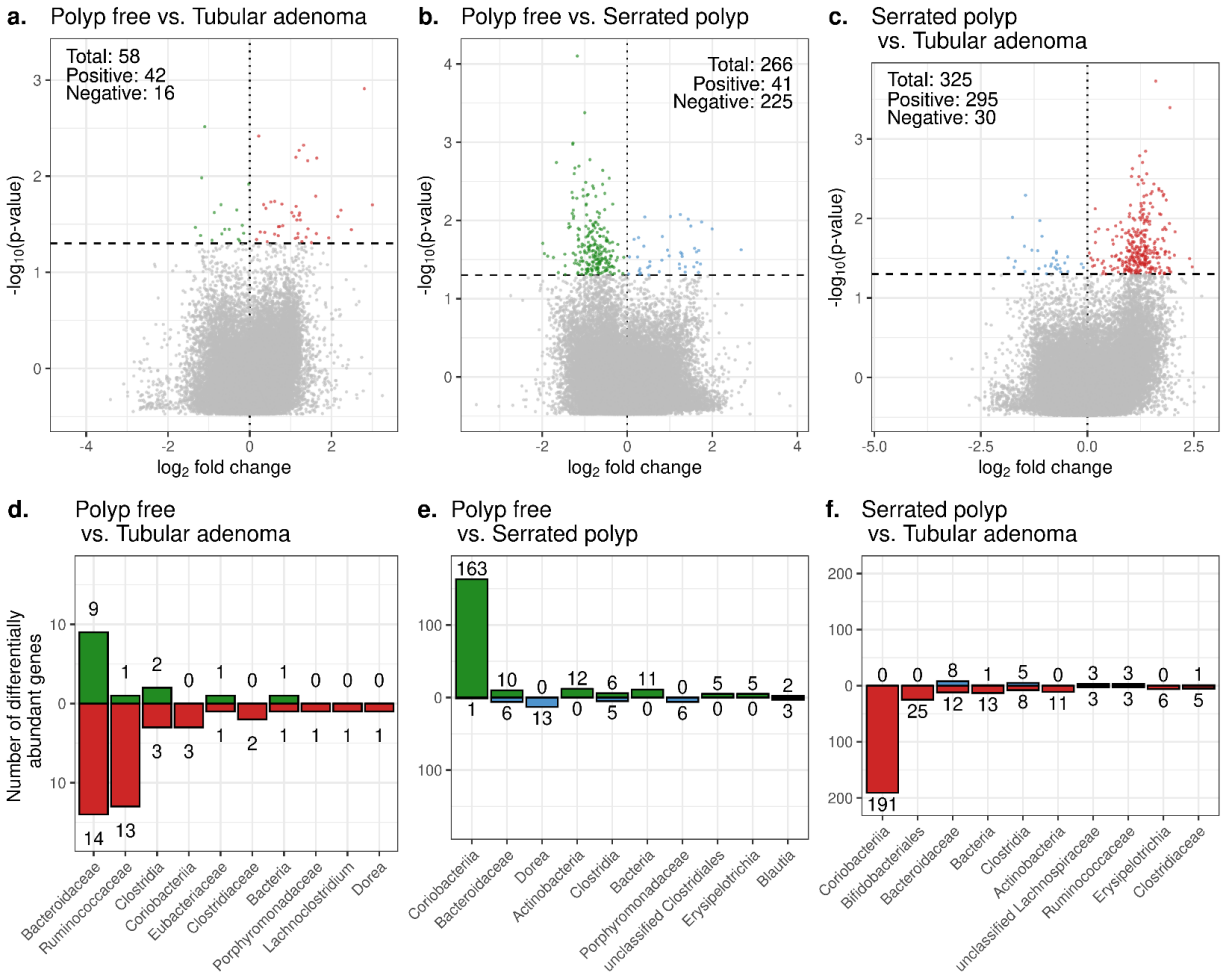
Supplementary Figure 1.7: A heatmap displaying the z-scores of microbial pathways from the second sample set. Samples are clustered by sample type and subject type. Within sample type, yellow represents mucosal aspirates, brown represents fecal samples, and purple represents lavage aspirates. Within subject type, green represents polyp free samples, blue represents serrated polyp samples, and red represents tubular adenoma samples. A total 507 pathways were identified.



Supplementary Figure 1.8: A receiver operating characteristic (ROC) curve illustrating the true positive rate (Sensitivity, y-axis) versus the false positive rate (Specificity, x-axis) produced by Random Forest classification of functional pathways in second sample set mucosal aspirates. The area under the curve (AUC) value for each Random Forest is displayed with the 90% confidence interval.

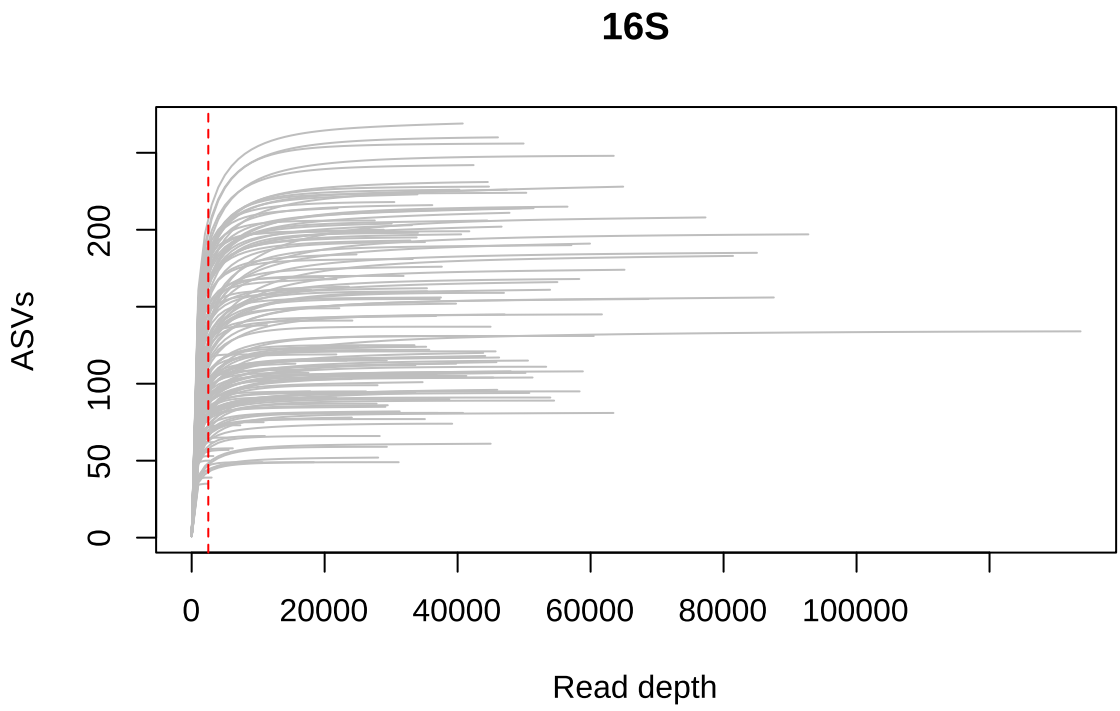


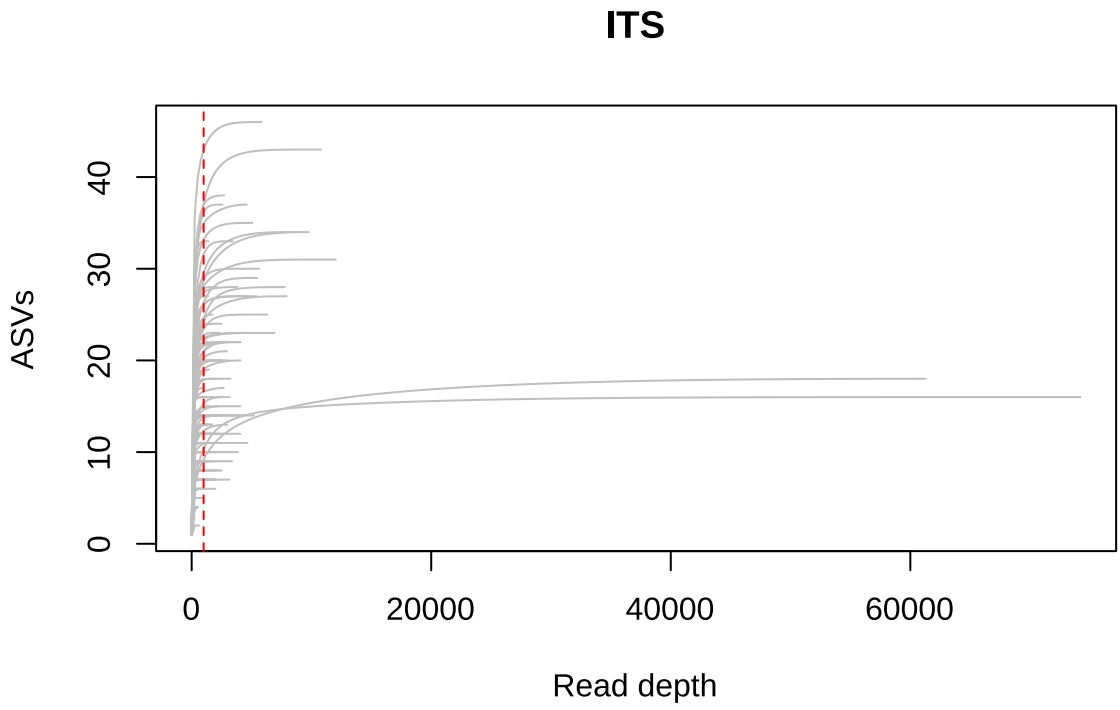
Supplementary Figure 1.9: **A-C)** The top ten variables of importance for each pairwise random forest classification of functional pathways in second sample set mucosal aspirates. Variables are sorted by their mean decrease in accuracy, with larger means contributing greater to Random Forest performance. **D)** Box plots displaying the functional pathway abundances (in reads per million) of the top variables of importance as determined by Random Forest. Each point is one sample, with multiple samples per individual. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range.



Supplementary Figure 1.10: **A-C**) Volcano plots illustrating the differentially abundant microbial genes within mucosal aspirate samples from the second sample set before FDR correction (Kruskal-Wallis: $p < 0.05$). The horizontal and vertical lines denote a significance threshold of $p = 0.05$, and zero $\log_2$fold change, respectively. Points are colored to denote the subject type in which the gene was more abundant, with green referring to genes more abundant in polyp-free samples, red for tubular adenomas, and blue for serrated polyps. The number of total, negative fold-change, and positive-fold change genes with an unadjusted p-value < 0.05 are displayed within each graph. **D-F**) The number of

differentially abundant genes per taxon for each subject type comparison. Only the top ten taxa with the most differentially abundant genes are shown.


16S
ASVs
0
50
100
200
0
20000
40000
60000
80000
100000
Read depth


ITS
ASVs
0
10
20
30
40
0
20000
40000
60000
Read depth

Supplementary Figure 1.11: Rarefaction curves of 16S (top) and ITS (bottom) amplicons from the first sample set. The x-axis is the read depth of each sample, with each line representing one sample. The y-axis is the number of unique ASVs per sample. The dotted red line represents the minimum required read depth for analysis inclusion. For 16S sequencing, this was determined at 2,500 high-quality, taxonomically annotated sequences, and for ITS it was 1,000 high-quality, taxonomically annotated sequences.

	Polyp-free	TA-bearing	SP-bearing (HPP/SSP)	Unknown or Other	Total
Mucosal brushes [On-polyp]	197 [0]	168 [58]	112 (61 [18]/ 51 [17])	48 [14]	525
Mucosal aspirates	350	280	195 (111/84)	72	897
Lavage aspirates	159	135	93 (54/39)	36	423
Fecal samples	9	17	9 (6/3)	3	38
Total	715	600	409	159	1883

Supplementary Table 1.1: A table showing the number of samples collected. Across rows, the number of each sample type is listed. For mucosal brushes, the number within the bracket corresponds to the number of brush samples taken directly from polyp tissue (as opposed to brushing non-polyp tissue). Across columns, the subject type classification is given. The number of samples per hyperplastic polyps (HPP) and sessile serrated polyps (SSP) are denoted parenthetically for the SP-bearing category. Samples were collected from a total of 140 unique individuals.

16S	Polyp-free	TA-bearing	SP-bearing (HPP/SSP)	Unknown or Other	Total
Mucosal brushes [On-polyp]	12 [0]	34 [11]	18 (8 [2]/10 [4])	0 [0]	64
Mucosal aspirates	17	24	11 (5/7)	0	52
Lavage aspirates	10	13	8 (3/5)	0	31
Fecal samples	0	0	0	0	0
Total	39	71	37	0	147

Supplementary Table 1.2: A table showing the number of samples with high quality sequencing reads in sample set 1, using 16S amplicon sequencing. Across rows, the number of each sample type is listed. For mucosal brushes, the number within the bracket corresponds to the number of brush samples taken directly from polyp tissue (as opposed to brushing non-polyp tissue). Across columns, the subject type classification is given. The number of samples per hyperplastic polyps (HPP) and sessile serrated polyps (SSP) are denoted parenthetically for the SP-bearing category. A total of 38 unique individuals were represented in this data.

ITS	Polyp-free	TA-bearing	SP-bearing (HPP/SSP)	Unknown or Other	Total
Mucosal brushes [On-polyp]	12 [0]	20 [7]	9 (1 [0]/8 [1])	0	41
Mucosal aspirates	13	17	11 (2/9)	0	41
Lavage aspirates	7	4	5 (1/4)	0	16
Fecal samples	0	0	0	0	0
Total	32	41	25	0	98

Supplementary Table 1.3: A table showing the number of samples with high quality sequencing reads in sample set 1, using ITS amplicon sequencing. Across rows, the number of each sample type is listed. For mucosal brushes, the number within the bracket corresponds to the number of brush samples taken directly from polyp tissue (as opposed to brushing non-polyp tissue). Across columns, the subject type classification is given. The number of samples per hyperplastic polyps (HPP) and sessile serrated polyps (SSP) are denoted parenthetically for the SP-bearing category. A total of 34 unique individuals were represented in this data.

WGS	Polyp-free	TA-bearing	SP-bearing (HPP/SSP)	Unknown or Other	Total
Mucosal brushes [On-polyp]	0	0	0	0	0
Mucosal aspirates	64	47	45 (24/17)	23	179
Lavage aspirates	5	11	4 (2/2)	1	21
Fecal samples	9	17	9 (6/3)	3	38
Total	78	75	58	27	238

Supplementary Table 1.4: A table showing the number of samples with high quality sequencing reads in sample set 2, using whole-genome shotgun sequencing. The number of each sample type is listed across rows. Across columns, the subject type classification is given. Additionally, the number of samples per hyperplastic polyps (HPP) and sessile serrated polyps (SSP) are denoted parenthetically for the SP-bearing category. A total of 117 unique individuals were represented in this data.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
BMI	1	1.41	1.40	9.83	0.03	0.001
AGE	1	0.98	0.98	6.88	0.02	0.001
ETHNICITY	4	5.18	1.29	9.07	0.10	0.001
SEX	1	1.29	1.29	9.00	0.02	0.001
SUBJECT TYPE	2	2.26	1.13	7.92	0.04	0.001
SUBJECT TYPE: INDIVIDUAL	28	27.15	0.97	6.79	0.51	0.001
SUBJECT TYPE: INDIVIDUAL: SAMPLE TYPE	63	7.96	0.13	0.88	0.15	0.992
RESIDUALS	46	6.57	0.14		0.13	
TOTAL	146	50.46			1.00	

PERMANOVA formula: 16S_ASV_table ~ BMI + Age + Ethnicity + Sex + Subject Type / Individual / Sample Type, strata = Plate

Supplementary Table 1.5: PERMANOVA analysis of brushes, mucosal aspirates, and lavage aspirates from the first sample set using 16S sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes polyp-free, tubular adenoma-bearing, and serrated polyp-bearing samples. Individuals are nested within subject type, and sample type is nested within the individual.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
BMI	1	0.53	0.53	1.26	0.01	0.066
AGE	1	0.43	0.43	1.03	0.01	0.341
SEX	1	0.44	0.44	1.05	0.01	0.297
ETHNICITY	3	1.40	0.47	1.12	0.03	0.108
SUBJECT TYPE	2	0.91	0.46	1.10	0.02	0.204
SUBJECT TYPE: INDIVIDUAL	24	11.77	0.49	1.17	0.28	0.003
SUBJECT TYPE: INDIVIDUAL: SAMPLE TYPE	38	16.03	0.42	1.01	0.38	0.361
RESIDUALS	25	10.43	0.42		0.26	
TOTAL	95	41.93			1.00	

PERMANOVA formula: ITS_ASV_table ~ BMI + Age + Sex + Ethnicity + Subject Type / Individual / Sample Type

Supplementary Table 1.6: PERMANOVA analysis of brushes, mucosal aspirates, and lavage aspirates from the first sample set using ITS sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes polyp-free, tubular adenoma-bearing, and serrated polyp-bearing samples. Individuals are nested within subject type, and sample type is nested within the individual.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
BMI	1	0.72	0.72	10.24	0.01	0.001
AGE	1	0.67	0.67	9.45	0.01	0.001
ETHNICITY	4	2.66	0.70	9.39	0.04	0.001
SEX	1	0.71	0.71	10.02	0.01	0.001
SUBJECT TYPE	2	1.43	0.72	10.07	0.02	0.001
SUBJECT TYPE: INDIVIDUAL	86	53.79	0.63	8.83	0.72	0.001
SUBJECT TYPE: INDIVIDUAL: SAMPLE TYPE	37	10.97	0.30	4.19	0.15	0.001
RESIDUALS	56	3.96	0.07		0.04	
TOTAL	188	74.94			1.00	

PERMANOVA formula: OTU_table ~ BMI + Age + Sex + Ethnicity + Subject Type / Individual / Sample Type, strata = Plate

Supplementary Table 1.7: PERMANOVA analysis of mucosal aspirates, lavage aspirates, and fecal samples from the second sample set using shotgun sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes polyp-free, tubular adenoma-bearing, and serrated polyp-bearing samples. Individuals are nested within subject type, and sample type is nested within the individual.

taxa_id	W	detected_0.9	detected_0.8	detected_0.7	detected_0.6
UBA1381.sp.	149	TRUE	TRUE	TRUE	TRUE
Ruminococcus.torques	147	TRUE	TRUE	TRUE	TRUE
Clostridium.ramosum	146	TRUE	TRUE	TRUE	TRUE
D16.sp..2	146	TRUE	TRUE	TRUE	TRUE
Ruminococcus.gnavus	145	TRUE	TRUE	TRUE	TRUE
Oscillibacter.sp.	145	TRUE	TRUE	TRUE	TRUE
Bacteroides.fragilis	143	TRUE	TRUE	TRUE	TRUE
Lachnospiraceae.sp..12	141	TRUE	TRUE	TRUE	TRUE
Dorea.formicigenerans	141	TRUE	TRUE	TRUE	TRUE
DTU089.HGM12760	141	TRUE	TRUE	TRUE	TRUE
D16.sp.	141	TRUE	TRUE	TRUE	TRUE
Ruminococcus.bicirculans	141	TRUE	TRUE	TRUE	TRUE
Clostridium.leptum	139	TRUE	TRUE	TRUE	TRUE
Eubacterium.HGM12316	136	FALSE	TRUE	TRUE	TRUE
Roseburia.intestinalis	136	FALSE	TRUE	TRUE	TRUE
Alistipes.sp..3	135	FALSE	TRUE	TRUE	TRUE
Coprococcus.catus	133	FALSE	TRUE	TRUE	TRUE
Ruminococcus.lactaris	131	FALSE	TRUE	TRUE	TRUE
Dorea.sp..2	130	FALSE	TRUE	TRUE	TRUE
Lachnospira.sp..2	130	FALSE	TRUE	TRUE	TRUE
Eggerthella.lenta	129	FALSE	TRUE	TRUE	TRUE
Flavonifractor.plautii	128	FALSE	TRUE	TRUE	TRUE
Eubacterium.sp..9	127	FALSE	TRUE	TRUE	TRUE
Lachnospira.pectinoschiza	124	FALSE	TRUE	TRUE	TRUE
Tyzzerella.sp..1	123	FALSE	FALSE	TRUE	TRUE
Faecalibacterium.HGM13278	123	FALSE	FALSE	TRUE	TRUE
Eubacterium.sp..6	121	FALSE	FALSE	TRUE	TRUE
Escherichia.coli	121	FALSE	FALSE	TRUE	TRUE
Lachnospiraceae.HGM11862	120	FALSE	FALSE	TRUE	TRUE
Erysipelatoclostridium.sp..2	119	FALSE	FALSE	TRUE	TRUE
Eubacterium.sp..15	119	FALSE	FALSE	TRUE	TRUE
Clostridium.bartlettii	119	FALSE	FALSE	TRUE	TRUE
Coprococcus.comes	116	FALSE	FALSE	TRUE	TRUE
DTU089.HGM12731	116	FALSE	FALSE	TRUE	TRUE
Faecalibacterium.sp..6	116	FALSE	FALSE	TRUE	TRUE
Bilophila.wadsworthia	115	FALSE	FALSE	TRUE	TRUE
Parabacteroides.distasonis	114	FALSE	FALSE	TRUE	TRUE

Intestinimonas.butyriciproducens	114	FALSE	FALSE	TRUE	TRUE
ER4.sp.	112	FALSE	FALSE	TRUE	TRUE
Dorea.longicatena.1	111	FALSE	FALSE	TRUE	TRUE
Clostridium.glycyrrhizinilyticum	111	FALSE	FALSE	TRUE	TRUE
UBA7182.HGM12585	109	FALSE	FALSE	TRUE	TRUE

Supplementary Table 1.8: Table of differentially abundant OTUs across fecal samples and mucosal aspirates from the second sample set using shotgun sequencing. Significance testing was performed using ANCOM2 (FDR < 0.05), adjusting for repeated measurements. "Detected 0.7" means that the microbe was differentially abundant in 70% of comparisons, which is the minimum for a microbe to be considered differentially abundant between categories.

taxa_id	W	detected_0.9	detected_0.8	detected_0.7	detected_0.6
Oscillibacter.sp.	95	TRUE	TRUE	TRUE	TRUE
Lachnospiraceae.sp..12	94	FALSE	TRUE	TRUE	TRUE
Ruminococcus.torques	84	FALSE	FALSE	TRUE	TRUE
Dorea.formicigenerans	83	FALSE	FALSE	TRUE	TRUE
Ruminococcus.bicirculans	76	FALSE	FALSE	TRUE	TRUE

Supplementary Table 1.9: Table of differentially abundant OTUs across fecal samples and lavage aspirates from the second sample set using shotgun sequencing. Significance testing was performed using ANCOM2 (FDR < 0.05), adjusting for repeated measurements. "Detected 0.7" means that the microbe was differentially abundant in 70% of comparisons, which is the minimum for a microbe to be considered differentially abundant between categories.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
COLON LOCATION	1	0.74	0.74	3.19	0.15	0.015
SUBJECT TYPE	1	0.42	0.42	1.83	0.08	0.124
SUBJECT TYPE: INDIVIDUAL	4	2.77	0.69	3.00	0.55	0.020
SUBJECT TYPE: INDIVIDUAL: TISSUE SITE	6	0.92	0.15	0.66	0.18	0.875
RESIDUALS	1	0.23	0.23		0.04	
TOTAL	13	5.07			1.00	

PERMANOVA formula: 16S_brushes_ASV_table ~ Colon location + Subject type / Individual / Tissue site

Supplementary Table 1.10: PERMANOVA analysis of brushes from the first sample set using 16S sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes tubular adenoma-bearing and serrated polyp-bearing samples. Tissue site includes polyp and healthy opposite wall brushes. Individuals are nested within subject type, and tissue site is nested within the individual.

A. Polyp-free vs. tubular adenoma mucosal aspirates

Factor	DoF	SoS	MS	F MODEL	R2	P-VAL
BMI	1	0.932049	0.932049	14.28858	0.02497	0.001
Age	1	0.510753	0.510753	7.829994	0.013683	0.001
Ethnicity	3	2.093075	0.697692	10.69581	0.056075	0.001
Sex	1	0.826017	0.826017	12.66306	0.02213	0.001
Colon Location	1	0.096297	0.096297	1.476259	0.00258	0.04
Prep Type	2	1.129164	0.564582	8.6552	0.030251	0.001
Subject Type	1	0.731753	0.731753	11.21797	0.019604	0.001
Subject Type: Patient	48	28.46318	0.592983	9.090592	0.762551	0.001
Residuals	39	2.543985	0.06523		0.068155	
Total	97	37.32627			1	

B. Polyp-free vs. serrated polyp mucosal aspirates

Factor	DoF	SoS	MS	F MODEL	R2	P-VAL
BMI	1	0.514723	0.514723	7.280071	0.014123	0.001
Age	1	0.610964	0.610964	8.641267	0.016763	0.001
Ethnicity	3	2.448353	0.816118	11.5429	0.067176	0.001
Sex	1	0.669934	0.669934	9.475329	0.018381	0.001
Colon Location	1	0.159277	0.159277	2.252762	0.00437	0.003
Prep Type	2	1.39528	0.69764	9.867187	0.038283	0.001
Subject Type	1	0.572898	0.572898	8.102883	0.015719	0.001
Subject Type: Patient	44	27.38873	0.622471	8.804025	0.75147	0.001
Residuals	38	2.686714	0.070703		0.073716	
Total	92	36.44687			1	

C. Tubular adenoma vs. serrated polyp mucosal aspirates

Factor	DoF	SoS	MS	F MODEL	R2	P-VAL
BMI	1	0.785529	0.785529	10.16161	0.023108	0.001
Age	1	0.848928	0.848928	10.98174	0.024973	0.001
Ethnicity	3	1.881841	0.62728	8.114507	0.055357	0.001
Sex	1	0.690819	0.690819	8.936442	0.020322	0.001
Colon Location	1	0.143375	0.143375	1.854707	0.004218	0.009
Prep Type	2	1.163965	0.581982	7.528535	0.03424	0.001
Subject Type	1	0.932052	0.932052	12.05705	0.027418	0.001
Subject Type: Patient	46	25.07423	0.545092	7.05132	0.737598	0.001
Residuals	32	2.473713	0.077304		0.072768	
Total	88	33.99446			1	

PERMANOVA formulas: Aspirate_OTU_table ~ BMI + Age + Ethnicity + Sex + Colon Location +Prep Type + Subject Type / Patient, strata = Plate

Supplementary Table 1.11: Pairwise PERMANOVA analysis of lavage aspirates from the second sample set using shotgun sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Prep type refers to the laxative used during colonoscopy prep. Table 11A compares polyp-free and tubular adenoma subject types, table 11B compares polyp-free and serrated polyp subject types, and table 11C compares tubular adenoma and serrated polyp subject types.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
SUBJECT TYPE	2	0.86	0.43	1.20	0.14	0.114
PREP TYPE	2	0.82	0.41	1.15	0.13	0.151
AGE	1	0.46	0.46	1.30	0.07	0.074
BMI	1	0.41	0.41	1.16	0.07	0.310
SEX	1	0.35	0.35	0.98	0.06	0.553
EHTNICITY	2	0.86	0.43	1.21	0.14	0.081
RESIDUALS	7	2.50	0.36		0.39	
TOTAL	16	6.27			1.00	

PERMANOVA formula: Lavage_OTU_table ~ Subject type + Prep type + Age + BMI + Sex + Ethnicity, strata = Plate

Supplementary Table 1.12: PERMANOVA analysis of lavage aspirates from the second sample set using shotgun sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes polyp-free, tubular adenoma-bearing and serrated polyp-bearing samples. Prep type refers to the laxative used during colonoscopy prep.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
SUBJECT TYPE	2	0.83	0.42	1.11	0.07	0.168
AGE	1	0.40	0.40	1.07	0.03	0.309
BMI	1	0.43	0.43	1.13	0.04	0.201
SEX	1	0.38	0.38	1.03	0.03	0.402
ETHNICITY	3	1.03	0.34	0.92	0.09	0.778
RESIDUALS	23	8.62	0.37		0.74	
TOTAL	31	11.70			1.00	

PERMANOVA formula: Fecal_OTU_table ~ Subject type + Age + BMI + Sex + Ethnicity, strata = Plate

Supplementary Table 1.13: PERMANOVA analysis of fecal samples from the second sample set using shotgun sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes polyp-free, tubular adenoma-bearing and serrated polyp-bearing samples. Prep type refers to the solution used during colonoscopy prep.

FACTOR	DoF	SoS	MS	F MODEL	R^2	P-VAL
BMI	1	0.004	0.004	19.36	0.015	0.001
AGE	1	0.002	0.002	8.76	0.007	0.001
ETHNICITY	4	0.013	0.003	14.26	0.043	0.001
SEX	1	0.001	0.001	4.28	0.003	0.012
SUBJECT TYPE	2	0.004	0.002	8.74	0.013	0.001
SUBJECT TYPE: INDIVIDUAL	86	0.221	0.003	11.73	0.768	0.001
SUBJECT TYPE: INDIVIDUAL: SAMPLE TYPE	38	0.031	0.001	3.72	0.108	0.001
RESIDUALS	57	0.013	0.001		0.043	
TOTAL	190	0.289			1.000	

PERMANOVA formula: Gene_table ~ BMI + Age + Ethnicity + Sex + Subject type / Individual / Sample type, strata = Plate

Supplementary Table 1.14: PERMANOVA analysis of functional genes within mucosal aspirates, lavage aspirates, and fecal samples from the second sample set using shotgun sequencing. The distance matrix method used was Bray-Curtis dissimilarity. Subject type includes polyp-free, tubular adenoma-bearing, and serrated polyp-bearing samples. Individuals are nested within subject type, and sample type is nested within the individual.

CHAPTER 2

Characterizing the microbiome of patients with myeloproliferative neoplasms during a Mediterranean diet intervention

Authors: Julio Avelar-Barragan, Laura F. Mendez Luque, Jenny Nguyen, Hellen Nguyen, Andrew Odegaard, Angela G. Fleischman, Katrine L. Whiteson

ABSTRACT

Myeloproliferative neoplasms (MPN) are a class of hematological malignancies which result in the overproduction of myeloid lineage cells. These malignancies result in increased cytokine production and inflammation, which correlate with worsened symptom burden and prognosis. Other than bone marrow transplantation, there is no cure for myeloproliferative neoplasms. As such, treatments focus on reducing thrombotic risk, inflammation, and symptom burden. Because current pharmacological treatments carry significant side-effects, there is a need to explore low-risk therapies. One alternative is the Mediterranean diet, which is rich in anti-inflammatory foods, reduces inflammatory biomarkers, and beneficially alters the gut microbiome. Here, we performed a 15-week clinical trial of 28 individuals with MPN who were randomized to dietary counseling based on either a Mediterranean diet or the standard U.S. Guidelines for Americans. Our primary objective was to determine if MPN patients could adopt a Mediterranean eating style with dietician counseling. As exploratory endpoints, we investigated the impact of diet and inflammation on the gut microbiome. Using shotgun metagenomic sequencing, we found that microbiome diversity and composition was stable throughout the study duration in both cohorts. Furthermore, we discovered significant alterations in the microbiomes between different MPN subtypes, such as increased beta-dispersion in subjects with

myelofibrosis. Lastly, we found several significant correlations between the microbiome and cytokines. Together, this study provides insight into the interaction between diet, inflammation, and the gut microbiome.

INTRODUCTION

Myeloproliferative neoplasms (MPN) are a group of hematological malignancies defined by somatic mutations which activate JAK/STAT signaling in hematopoietic stem cells. [1,2] This results in an overproduction of myeloid lineage cells. Clinically, MPNs are divided into three clinical phenotypes: polycythemia vera (PV), essential thrombocythemia (ET), and myelofibrosis (MF). PV is characterized by an elevated red blood cell mass. Elevations in platelets and white blood cells are also common. Subjects with ET have elevated platelets but rarely have increased red or white blood cells. MF is characterized by reticulin fibrosis in the bone marrow, and often cytopenia. MF can develop from a "burn out" phase following PV or ET, termed post-PV or post-ET MF, or without a preceding diagnosis of PV or ET, termed primary myelofibrosis.

One feature of MPN is increased inflammatory cytokine abundance, which correlates with worsened symptom burden and disease prognosis. [3,4] MPN symptom burden can be severe, and many individuals experience fatigue, early satiety, abdominal discomfort, night sweats, pruritus, bone pain, fever, and unintentional weight loss. Other than bone marrow transplantation, there is no cure for MPN. MPN management focuses on reducing thrombotic risk and alleviating symptom burden. Current pharmacological treatments for MPN includes JAK inhibitors, such as ruxolitinib, but these often carry significant side effects such as immunosuppression. [5] Consequently, there is a need to explore low-risk alternatives for MPN management.

One method to non-pharmacologically manage MPN is through the consumption of a Mediterranean (MED) diet, consisting of extra virgin olive oil, fruits, vegetables, whole grains, legumes, fish, nuts, and seeds. A MED diet has been shown to reduce inflammation by lowering C reactive protein and IL-6 levels and is associated with reduced obesity, cardiovascular disease, and cancer risk. [6-9] Adherence to a MED diet has been found to alter the gut microbiome, which is the collection of bacteria, fungus, viruses, and other microorganisms living within the large intestine. [10-15] Mechanistically, this may occur because of the fermentation of dietary fiber and unsaturated fat by the gut microbiota to produce anti-inflammatory metabolites. [16] However, it remains to be seen whether a MED diet can be strategically used to manipulate the gut microbiota to promote health by reducing inflammation in MPN.

We performed a randomized clinical trial to investigate whether registered dietician counseling can alter the eating pattern of individuals with MPN toward a Mediterranean style. Subjects were randomly assigned to either MED diet counseling supplemented with complementary extra virgin olive oil or counseling following the standard US Guidelines for Americans (USDA) supplemented with grocery certificates. The study length was 15 weeks, consisting of 2 weeks pre-intervention observation, 10 weeks of active dietary counseling, and 3 weeks of post-counseling follow-up. As a key exploratory endpoint, we investigated if a MED diet could produce a microbiome-mediated reduction in inflammation. Blood and stool samples were collected to measure cytokine levels and assess gut microbiome composition, respectively. Survey data was collected to assess the feasibility of a MED diet intervention among MPN patients and symptom burden was tracked using the MPN Symptom Assessment Form (MPN-SAF). In a companion manuscript we describe the

relationship between MED diet adherence, symptom burden, and cytokine concentrations. In this manuscript, we detail the association of the gut microbiota with diet, MPN subtype, and cytokine concentrations.

RESULTS

Cohort description and study synopsis:

Twenty-eight subjects with MPN were recruited for this study (Figure 2.1). The MED cohort had 15 individuals, while the USDA had 13 individuals. Within the MED cohort, 3 subjects had ET, 4 had MF, and 8 had PV. Within the USDA cohort, 3 subjects had ET, 4 had MF, and 6 had PV. The median age for the MED cohort was 59 +/- 14.5 (σ) years, while the median age for the USDA cohort was 61 +/- 14 (σ) years. Both groups had 10 females each, with 5 and 3 males in the MED and USDA cohorts, respectively. The study took place over 15 weeks and had an active intervention period from weeks 3-12. Baseline blood and stool samples were collected at week 1, followed by additional sampling during the active intervention at weeks 6 and 9. Follow-up samples were also taken after the intervention's end at week 15. Throughout the study, six unannounced surveys and 24-hour food recalls (ASA24) were collected to measure diet compliance, and symptom burden was assessed using the MPN-Symptom Assessment Form (MPN-SAF), which grades the 10 most clinically relevant symptoms of MPN patients. [57] Table 2.1 provides a detailed description of each subject's characteristics.

Gut microbiome diversity and composition is stable during Mediterranean diet intervention:

We began our investigation by examining how a Mediterranean Diet (MED) impacts gut microbiome diversity. Analysis of species richness estimates using a linear mixed-

effects model (LME) demonstrated that USDA and MED groups did not significantly differ over time after accounting for pre-intervention differences (LME p-value = 0.48, Figure 2.2A). Analyses of species evenness estimates showed no differences between diet groups either (LME p-value = 0.65, Figure 2.2B). Sub-setting species richness and evenness comparisons to include only samples from participants highly adherent to a Mediterranean style eating pattern and those least adherent to a Mediterranean style eating pattern during the intervention did also not reveal significant differences (LME richness p-value = 0.48, LME evenness p-value = 0.73).

Next, we examined the microbial composition, or beta-diversity, of our samples. Species composition analysis using non-metric multidimensional scaling (NMDS) and permutational multivariate analysis of variance (PERMANOVA) showed that there were significant differences associated with MED and USDA groups pre-intervention (Figure 2.2C; PERMANOVA $R^2 = 0.057$, p-value = 0.046, Supplementary Table 2.1). Therefore, we stratified our PERMANOVA analysis to investigate whether gut microbiome composition changed over time within each individual. This produced non-significant results, suggesting that microbiome composition was stable over the duration of the study (Figure 2.2D; PERMANOVA $R^2 = 0.007$, p-value = 0.76, Supplementary Table 2.2). Next, PERMANOVA was performed on each MPN subtype to examine whether a specific subtype responded to the diet intervention more than others. No changes were detected in ET (PERMANOVA $R^2 = 0.046$, p-value = 0.63), MF (PERMANOVA $R^2 = 0.026$, p-value = 0.60), or PV (PERMANOVA $R^2 = 0.016$, p-value = 0.77) subtypes over time (Supplementary Table 2.3). Consequently, we did not find any differentially abundant microbes between MED and USDA groups after adjusting for pre-existing compositional differences.

Characterization of the functional metagenome demonstrated no significant differences between diets as measured by microbial gene richness (LME p-value = 0.65, Supplementary Figure 2.1A) and gene evenness (LME p-value = 0.19, Supplementary Figure 2.1B) after accounting for pre-intervention differences. PERMANOVA analysis indicated that there were no significant changes over time within each individual (PERMANOVA R^2 = 0.009, p-value = 0.29, Supplementary Table 2.4). Thus, no differentially abundant genes were found between MED and USDA groups.

Individuals with myelofibrosis have reduced microbial diversity and altered composition:

Previous research has demonstrated significant differences in the microbiomes associated with healthy individuals and those with MPN. [17] Therefore, we further characterized the microbiome between PV, ET, and MF subtypes. Using species richness estimates, we observed a significant reduction in the number of unique microbes when comparing individuals with MF to PV (Linear mixed-effects p-value = 0.028, Supplementary Figure 2.2A), and a non-significant reduction when comparing MF to ET (Linear mixed-effects p-value = 0.056, Figure 2.2A). Species abundance distribution, or evenness, was also reduced in MF, but was not significant compared to PV (LME p-value = 0.12) and ET subtypes (LME p-value = 0.47, Figure 2.2B).

With respect to beta-diversity, samples from MF were more dissimilar from each other, resulting in a trend towards increased beta-dispersion when compared to ET (LME p-value = 0.089) and PV (LME p-value = 0.056, Supplementary Figure 2.2C and Figure 2.3A). Conversely, PV and ET samples tended to cluster together (LME p-value = 0.895, Supplementary Figure 2.2C and Figure 2.3A). PERMANOVA demonstrated a significant

association with microbial composition and MPN subtype, explaining approximately 6.1% of observed variance (PERMANOVA p-value = 0.001, Supplementary Table 2.5). Microbiomes were largely personalized, with the individual of origin significantly explaining about 54% of the variance observed in the microbiome (PERMANOVA p-value = 0.001, Supplementary Table 2.5). *Faecalibacterium prausnitzii* was depleted in subjects with MF when compared to those with PV and ET (ANCOM2 $p < 0.05$, Figure 2.3B). Microbes correlated with *F. prausnitzii* abundance included *Ruminococcus torques, Coprococcus catus, Agathobaculum butyriciproducens, Ruminococcus gnavus, Clostridium bolteae,* and *Blautia sp.* CAG-257 (Figure 2.3C).

Within our functional metagenomes, we detected a significant reduction in the number of unique microbial genes within MF subjects when compared to PV (LME p-value = 0.016, Supplementary Figure 2.3A), but not ET (Linear mixed-effects p-value = 0.244, Supplementary Figure 2.3A). There was no significant difference in the gene evenness among MPN subtypes (Supplementary Figure 2.3B). NMDS ordination demonstrated that the functional metagenome compositions of MF samples tended to be more disparate from each other when compared to PV (LME p-value = 0.34) and ET (LME p-value = 0.19, Supplementary Figure 2.3C-D). Additionally, MPN subtype significantly explained about 6.7% of the variance observed in functional metagenome composition (PERMANOVA p-value = 0.001, Supplementary Table 2.6), while the individual of origin was associated with about 54% of the variance (PERMANOVA p-value = 0.001, Supplementary Table 2.6). Differential abundance analysis produced no significantly different genes between MPN subtypes after FDR correction.

Cytokine levels are correlated with microbiome diversity and composition:

After our subsequent analysis of MPN subtypes and their gut microbiomes, we next asked if the microbiome was associated with the levels of ten cytokines. Comparison of cytokine concentrations between MPN subtypes revealed a significant increase of TNFα and IL-12p70 in subjects with MF when compared to ET (Tukey's test; TNFα p-adj < 0.001 and IL-12p70 p-adj = 0.016) and PV (Tukey's test; TNFα p-adj = 0.002 and IL-12p70 p-adj = 0.022, Figure 2.4A). IL-6, IL-8, and IL-10 concentrations were elevated in subjects with MF but were not statistically significant (Figure 2.4A). Correlation of cytokines with microbial richness resulted in a negative correlation with TNFα (Spearman's ρ = -0.50, p-adj = 0.07, Figure 2.4B) and IL-12p70 (Spearman's ρ = -0.45, p-adj = 0.15, Figure 2.4C).

Next, we correlated cytokines with microbial abundances at the genus level, resulting in 34 significant correlations (Figure 2.4D). Notable correlations included associations with TNFα vs. *Flavonfractor* (Spearman's ρ = 0.39, p = 0.038), IL-12p70 vs. *Roseburia* (Spearman's ρ = -0.55, p = 0.002), and IL-8 vs. *Eubacterium* (Spearman's ρ = -0.41, p = 0.032, Supplementary Figure 2.4). Similarly, we compared cytokines with functional pathway abundances, producing 162 significant correlations (Supplementary Figure 2.5). Notable correlations included TNFα vs. 4-deoxy-L-threo-hex-4-enopyranuronate degradation (Spearman's ρ = -0.42, p = 0.038), TNFα vs. β-(1,4)-mannan degradation (Spearman's ρ = -0.48, p = 0.019), and IL-12p70 vs. GDP-mannose biosynthesis (Spearman's ρ = -0.59, p = 0.003, Supplementary Figure 2.6).

DISCUSSION

Our goal with this manuscript was to 1) assess whether a MED diet altered the gut microbiome of subjects with MPN and 2) to investigate the association between the gut microbiome and cytokines. In a separate manuscript (Mendez Luque, in preparation), we

describe the feasibility of a MED diet intervention in the MPN subject population, changes in macronutrients associated with the dietary intervention, and the interaction between diet adherence, symptom burden, and cytokine concentrations. Here, we found no significant changes in microbial diversity or composition associated with a MED diet intervention over a 10-week active dietary intervention period. Instead, we found the MPN subtype played a greater role in determining microbiome diversity and composition. Individuals with ET and PV had more similar microbial compositions, while those with MF were more disparate. Furthermore, a reduction of microbial diversity correlated with elevated TNFα and IL-12p70 concentrations in subjects with MF. These differences in cytokine concentrations were associated with the abundance of microbial genera and metabolic pathways, further establishing a role for the gut microbiome in inflammation and MPN.

With respect to diet-mediated changes in the microbiome, there are multiple explanations as to why the microbiomes of individuals remained stable throughout the dietary intervention. The first is intervention duration. Long term adherence to a Mediterranean diet has been demonstrated to reduce the incidence of cardiovascular disease, Alzheimer's disease, colorectal cancer, diabetes, and obesity, but it remains unclear how long individuals must adhere to the diet to manifest its benefits. [8,9] Studies performing MED diet interventions have ranged from 6 weeks to 7 years. [12-15,18-26] Of studies examining gut microbiome composition, the 6-week MED diet intervention performed by Marlow *et al* yielded no significant differences in gut microbiome composition or CRP levels in subjects with Crohn's disease. [26] Comparatively, Nagpal *et al* conducted a MED diet intervention in a non-human primate model over the course of 2.5 years, where a significant difference was

observed between macaques who consumed a Western diet versus a MED diet.[14] Although diet has been shown to rapidly alter the composition of the microbiome, our study and others suggest that longer dietary interventions are needed to detect changes in gut microbiome composition, especially with small sample sizes.[27]

Another consideration in the successful manipulation of gut microbiomes with diet is presence of specific microbial taxa, functions, or enterotypes. Stratification of microbiomes into enterotypes has revealed enterotype-specific predictors to dietary intervention response.[28] Klimenko *et al.* found that the strongest predictor of whether an individual would respond to a dietary intervention was the average number of genes per microbe.[28] A negative correlation between the average number of genes per microbe and alpha diversity was also found, suggesting that more diverse communities are formed by specialist microbes with fewer genes.[28] The microbiomes associated with industrialized countries, like the United States, often have reduced diversity and a higher abundance of *Bacteroides* when compared to non-industrialized countries.[29] Many *Bacteroides* are generalists, meaning they contain more genes and wider metabolic potentials than specialist taxa.[30] The predominance of generalist taxa has been known to contribute to microbiome stability.[31] Therefore, it is plausible that the microbiomes of industrialized individuals have evolved to resist perturbations, such as those caused by antibiotic usage or short-term diet changes. In one study, individuals with a higher ratio of *Prevotella* to *Bacteroides* lost more weight than those with a lower *Prevotella* to *Bacteroides* ratio while consuming a New Nordic Diet, suggesting that a higher abundance of generalists is associated with intervention outcome too.[32] Our samples contained a significant proportion of *Bacteroides*, so it is possible that the microbiomes of these individuals were

resistant to short-term dietary changes as reflected by the non-significant changes in diversity, composition, and function over time.

One final factor which could affect the stability of microbiomes is the strength of the dietary intervention. Due to differences in agriculture and food processing, a MED diet in United States is likely different than a MED diet in the Mediterranean. This can affect the number of antibiotic and prebiotic compounds found in each diet. One prebiotic component of the MED diet that can influence gut microbiome composition is extra virgin olive oil (EVOO). EVOO is rich in polyphenols and oleic acid, which have been demonstrated to have anti-oxidative and anti-inflammatory properties. [33,34] Over 90% of polyphenols are digested and metabolized in the colon by the gut microbiota. [35] Dietary supplementation of EVOO in humans has been shown to promote the growth of beneficial microbes like *Bifidobacterium* and lactic acid producing bacteria. [33,36] In rodent models, consumption of EVOO results in an increased abundance of *Bifidobacterium, Lactobacillus,* and *Clostridium.* [37,38]

The MED diet is also typically higher in dietary fiber when compared to a typical USDA diet. Dietary fiber is fermented by the gut microbiota to produce short-chain fatty acids, like acetate, propionate, and butyrate. Butyrate is critical for gut health, as it the primary source of energy for colonocytes and reduces inflammation by stimulating the production of T-regulatory cells and IL-10 producing cells. [39] In this study, we did not find any differences in the amount of butyrate-producing bacteria between diets. Instead, we saw a reduction of the butyrate-producing microbe, *F. prausnitzii,* in subjects with MF. We also noted significant positive correlations between *F. prausnitzii, Agathobaculum butyriciproducens,* and *Coprococcus catus* abundances. *A. butyriciproducens* is another

butyrate-producing microbe, while *C. catus* produces butyrate and propionate. [40] We also observed broader, community wide differences between subjects with ET, PV, and MF. Notably, the microbiome composition of MF subjects was more dissimilar to each other when compared to ET and PV. Our previous work comparing the gut microbiome composition of healthy and MPN subjects similarly showed that individuals with MF had increased beta dispersion when compared to ET and PV. [17] These results describe a phenomenon known as the 'Anna Karenina principal' for animal microbiomes, which states that stressors affect microbiomes in unpredictable ways, leading to increased community beta-dispersion. [41,42]

One likely stressor resulting in higher MF beta-dispersion is the increased concentration of pro-inflammatory cytokines. Inflammation has been known negatively affect the gut microbiome. Supporting this notion, we found that TNFα and IL-12p70 were significantly increased in MF subjects, which negatively correlated with species richness overall. We found IL-12p70 negatively correlated with the genus, *Roseburia*, which are butyrate producing microorganisms known to alleviate inflammation by promoting T-regulatory cell differentiation. [43,44] We also observed a significant negative correlation with *Eubacterium* and the pro-inflammatory cytokine, IL-8. *Eubacterium* also produce butyrate and have been shown to lessen inflammation by promoting IL-10 production. [43,45] TNFα and the 4-deoxy-L-threo-hex-4-enopyranuronate degradation pathway negatively correlated as well. This pathway plays a role in the degradation of uronic acids, such as apple pectin. β-(1,4)-mannan is another compound found in plant cell walls and the pathway for its degradation was found to be negatively correlated with TNFα.

Taken together, it is possible that the increased inflammation observed in individuals with MPN, particularly MF, is exacerbated by the lack of sufficient short-chain fatty acid production. The MED diet has been previously shown to promote the growth of *F. prausnitzii* specifically, therefore, future experiments could attempt restore microbial short-chain fatty acid production to reduce inflammation. When designing dietary interventions, however, special attention should be given to the intervention duration and the ability for existing gut microbes to use and respond to prebiotic compounds. This may ensure that the desired outcomes are achieved, allowing us to manipulate the gut microbiome to promote health and ameliorate disease.

METHODS

Recruitment of subjects:

Patients were recruited between October 2018 and September 2019. Participants were included if they were over the age of 18 with a previous diagnosis of a Philadelphia chromosome negative MPN (including PV, ET, MF), had an ECOG score of 2 or less, a life expectancy of greater than 20 weeks, had internet access with an email address, and could read and understand English. Any type of previous or current therapy was also allowed. Participants were excluded if they were pregnant or planning on becoming pregnant, lost more than 10 pounds or 10% of their body weight in the last 6 months, or were allergic to nuts and olive oil. Forty-seven participants were screened. Five did not meet the inclusion criteria, and an additional 11 subjects were excluded due to incomplete survey data during the observation period. Thirty-one subjects were randomly assigned to a diet, but 2 withdrew participation and one failed to provide sufficient survey data. This final number

of study participants was 28, with 15 belonging to the MED cohort and 13 belonging to the USDA cohort.

Collection of dietary intervention feasibility, adherence, and symptom burden data:

During the first week of the intervention period, each participant met individually with a dietician to learn about the central components of their assigned diet. In addition, there were follow-up dietician visits during weeks 5 and 7. Participants were emailed educational materials on their respective diet weekly during the 10-week active intervention period. Furthermore, participants in the MED cohort were given 750 mL of extra virgin olive oil and those in the USDA cohort were given a $10 USD grocery gift card at weeks 3 and 6. Throughout the study duration, participants were required to fill out 4 unannounced surveys given during weeks 1, 2, 3, 6, 9, 12, and 15. The first survey measured dietary intervention feasibility and asked, "how easy is it for you to follow this diet, with 1 being very easy to follow and 10 being very difficult to follow?" The second survey measured MED diet adherence. For this, the established 14-item Mediterranean diet adherence score (MEDAS) was used. [46] Adherence to a MED diet was defined as a "high" for the week if a score of >8/14 was obtained. Next, we asked subjects to complete 24-hour food recalls by using the Automated Self-Administered 24-hour Dietary Assessment Tool (ASA24). Lastly, symptom burden was assessed via the MPN symptom assessment form (MPN-SAF), which grades the 10 most clinically relevant MPN symptoms. [47] Surveys were administered through email.

Blood collection and cytokine measurements:

Peripheral blood was drawn on weeks 1, 3, 6, and 15 in tubes containing ethylenediaminetetraacetic acid (EDTA). Plasma was obtained by centrifuging 3-4 ml of

blood for 10 minutes at 2500 rpm, aliquoted, and was stored at -80°C. Frozen plasma was sent to Quanterix in Billerica, MA for analysis. A Human CorPlex 10 Cytokine Array kit #85-0329 (IL-12p70, IL-1B, IL-4, IL-5, IFNɣ, IL-6, IL-8, IL-22, TNFα, and IL-10) was used according to manufacturer's protocol and analyzed using A Quanterix SPX imager system on-site at Quanterix Headquarters in Billerica, MA.

Fecal sample collection:

To perform gut microbiome analysis, four stool samples were requested from each participant over the course of the 15-week trial. The samples were collected by the participants themselves using Zymo RNA/DNA shield fecal collection tubes (Cat. #R1101) during weeks 1, 6, 9, and 15. Samples were returned in person or by mail. In total, 103 samples were collected. Samples were stored at -80°C once returned.

DNA extraction:

Fecal samples stored in DNA/RNA shield were thawed on ice, vortexed to homogenize, then 1000 uL of fecal slurry was extracted using ZymoBiomics DNA Miniprep Kit (Cat. #D4300) according to the manufacturer's protocol. Bead lysis during the extraction was performed at 6.5 m/s for 5 minutes total (MPBio FastPrep-24).

Shotgun library preparation and sequencing:

Libraries for shotgun sequencing were prepared using the Illumina DNA prep kit (Cat. # 20018705), using an adapted low-volume protocol. [48] In summary, we reduced the amount of DNA used per sample to a maximum of 5 uL or 50 ng (whichever was reached first). Tagmentation was performed according the manufacturer's protocol, but volumes were reduced to 1 uL of bead-linked transposome and tagmentation buffer each. Next, 1.25 uL of 1 uM i5 and i7 indices were added to each sample each and annealed via polymerase

chain reaction using 10 uL of KAPA HiFi HotStart ReadyMix (Cat. # 7958935001). Afterwards, libraries were combined, size-selected, and cleaned using 56 and 14.4 uL of sample purification beads according to the low-volume protocol. Positive and negative sequencing controls were included during the library preparation using the ZymoBIOMICS Microbial Community DNA Standard (Cat. #D6305) and purified water, respectively. The quality of libraries was assessed with Quanti-iT PicoGreen dsDNA (Cat. #P7589) for quantity and Agilent Bioanalyzer High Sensitivity DNA Analysis (Cat. #5067-4626) for fragment size. Libraries were shipped overnight on dry ice to Novogene Corporation Inc. (Sacramento, CA) to be sequenced using Illumina's Hiseq 4000. An average of 2,819,107 +/- 670,543 (σ) paired-end reads per sample, 150 base-pairs long, were obtained.

OTU table generation:

Raw data was first cleaned to remove sequencing adapters and artifacts using the BBMap v38.79 script 'bbduk.sh' with the flag 'ref=adapters,artifacts'. [49] BBMap's 'demuxbyname.sh' was used to demultiplex sequences using the default parameters. Quality filtering of sequences was performed using PRINSEQ++ v1.2 with the following parameters: -trim_left 5 -trim_right 5 -min_len 100 -trim_qual_right 28 -min_qual_mean 25. [50] Quality checking was done with FastQC v0.11.8 on default parameters. This resulted a mean and standard deviation of 2,731,886 +/- 648,042 paired-end reads, respectively. Human-derived reads were removed using BowTie2 v2.4.5 using the default parameters and hg38 as the reference genome, which produced an average of 2,498,159 +/- 960,477 (σ) reads per sample. [51] Taxonomic assignment of the resulting sequences was assigned using MetaPhlAn v3.0.14 with default parameters and the CHOCOPhlAn v2019.01 database. [52]

Microbiome functional potential data generation:

Individual gene annotations were produced by first cross-assembling reads into contiguous sequences using MEGAHIT v1.1.1 with a minimum length of 2,500 base pairs and the flag '--k-list 31,41,51,61,71,81,91,101,111'. [53] Afterwards, open reading frame were assigned with Prodigal v2.6.3 and then annotated with eggNOG mapper v2.0 using the eggNOG v5.0 database. [54,55] Next, BowTie2 v2.4.5 was used to align samples to the annotated genes to obtain a table of per sample counts for each gene. Lastly, per sample gene counts were normalized to reads per kilobase per genome equivalent using MicrobeCensus v1.1.1 on default parameters. [56] For functional annotation of metabolic pathways, we ran our quality-filtered, unassembled reads through HUMAnN v3.0.1 using the default parameters and the UniRef90 v201901b database. [52] The 'humann_renorm_table' and 'humann_join_tables' scripts were used to create a pathway abundance table of normalized counts in copies per million.

Data analysis:

Data analysis of OTUs, genes, and pathways was performed in R v4.2.1. The first step was removing microbes or genes which contaminated our sequencing controls from all samples. The Vegan v2.6-2 package was used to calculate the following metrics: richness with the 'specnumber' function, evenness with the formula 'diversity(x, index = "Shannon") / $\log_{10}$(specnumber(x))', Bray-Curtis beta diversity with the 'vegdist' function, PERMANOVA with the 'adonis2' function, NMDS with the 'metaMDS' function, and beta-dispersion with the 'betadisper' function. Please see Supplementary Tables 1 – 6 for PERMANOVA formulas and parameters. Significance testing of richness, evenness, and beta dispersion was performed using linear-mixed effect models with the nlme v3.1-159

package (Pinheiro 2020). Significance testing of cytokine concentrations was done using an ANOVA and Tukey's post-hoc test with the 'aov' and 'TukeyHSD' functions. Spearman correlations were obtained using the 'cor.test' and 'rcorr' functions. Differential abundance of OTUs was determined with ANCOM v2.1 with the parameters: rand_formula = "~ 1 | Subject", p_adj_method = "none", alpha = 0.05. We were unable to perform ANCOM for gene and pathway abundances, therefore, we averaged abundances within each subject to eliminate repeated measurements and performed a Kruskal-Wallis test. When appropriate, multiple comparisons were corrected for using the 'p.adjust(x, method = "fdr")' function. All code, scripts, and parameters for data processing and analysis can be found at https://github.com/Javelarb/MPN_diet_intervention.

REFERENCES

1. Kralovics, R. et al. A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders. N Engl J Med 352, 1779–1790 (2005).
2. Nangalia, J. et al. Somatic CALR Mutations in Myeloproliferative Neoplasms with Nonmutated JAK2. N Engl J Med 369, 2391–2405 (2013).
3. Fleischman, A. G. Inflammation as a Driver of Clonal Evolution in Myeloproliferative Neoplasm. Mediators of Inflammation 2015, 1–6 (2015).
4. Craver, B., El Alaoui, K., Scherber, R. & Fleischman, A. The Critical Role of Inflammation in the Pathogenesis and Progression of Myeloid Malignancies. Cancers 10, 104 (2018).
5. Tefferi, A. JAK inhibitors for myeloproliferative neoplasms: clarifying facts from myths. Blood 119, 2721–2730 (2012).
6. Smidowicz, A. & Regula, J. Effect of Nutritional Status and Dietary Patterns on Human Serum C-Reactive Protein and Interleukin-6 Concentrations. Advances in Nutrition 6, 738–747 (2015).
7. Estruch, R. Anti-inflammatory effects of the Mediterranean diet: the experience of the PREDIMED study. Proc. Nutr. Soc. 69, 333–340 (2010).
8. Sofi, F., Abbate, R., Gensini, G. F. & Casini, A. Accruing evidence on benefits of adherence to the Mediterranean diet on health: an updated systematic review and meta-analysis. The American Journal of Clinical Nutrition 92, 1189–1196 (2010).
9. Gotsis, E. et al. Health Benefits of the Mediterranean Diet: An Update of Research Over the Last 5 Years. Angiology 66, 304–318 (2015).

10. De Filippis, F. et al. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut 65, 1812–1821 (2016).
11. Mitsou, E. K. et al. Adherence to the Mediterranean diet is associated with the gut microbiota pattern and gastrointestinal characteristics in an adult population. Br J Nutr 117, 1645–1655 (2017).
12. Ghosh, T. S. et al. Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-year dietary intervention across five European countries. Gut 69, 1218–1228 (2020).
13. Meslier, V. et al. Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. Gut 69, 1258–1268 (2020).
14. Nagpal, R. et al. Gut Microbiome Composition in Non-human Primates Consuming a Western or Mediterranean Diet. Front. Nutr. 5, 28 (2018).
15. Haro, C. et al. Two Healthy Diets Modulate Gut Microbial Community Improving Insulin Sensitivity in a Human Obese Population. The Journal of Clinical Endocrinology & Metabolism 101, 233–242 (2016).
16. Bailey, M. A. & Holscher, H. D. Microbiome-Mediated Effects of the Mediterranean Diet on Inflammation. Advances in Nutrition 9, 193–206 (2018).
17. Oliver, A. et al. Fecal Microbial Community Composition in Myeloproliferative Neoplasm Patients Is Associated with an Inflammatory State. Microbiol Spectr 10, e00032-22 (2022).

18. Shively, C. A. et al. Mediterranean versus Western Diet Effects on Caloric Intake, Obesity, Metabolism, and Hepatosteatosis in Nonhuman Primates: Mediterranean and Western Diet in Nonhuman Primates. Obesity 27, 777–784 (2019).
19. Garcia-Rios, A. et al. Beneficial effect of CLOCK gene polymorphism rs1801260 in combination with low-fat diet on insulin metabolism in the patients with metabolic syndrome. Chronobiology International 31, 401–408 (2014).
20. Kaaks, R. et al. Effects of dietary intervention on IGF-I and IGF-binding proteins, and related alterations in sex steroid metabolism: the Diet and Androgens (DIANA) Randomised Trial. Eur J Clin Nutr 57, 1079–1088 (2003).
21. Delgado-Lista, J. et al. CORonary Diet Intervention with Olive oil and cardiovascular PREVention study (the CORDIOPREV study): Rationale, methods, and baseline characteristics. American Heart Journal 177, 42–50 (2016).
22. Shai, I. et al. Weight Loss with a Low-Carbohydrate, Mediterranean, or Low-Fat Diet. N Engl J Med 359, 229–241 (2008).
23. Mekki, K., Bouzidi-bekada, N., Kaddous, A. & Bouchenak, M. Mediterranean diet improves dyslipidemia and biomarkers in chronic renal failure patients. Food & Funct. 1, 110 (2010).
24. Paniagua, J. A. et al. A MUFA-Rich Diet Improves Posprandial Glucose, Lipid and GLP-1 Responses in Insulin-Resistant Subjects. Journal of the American College of Nutrition 26, 434–444 (2007).
25. Salas-Salvadó, J. et al. Prevention of Diabetes With Mediterranean Diets: A Subgroup Analysis of a Randomized Trial. Ann Intern Med 160, 1–10 (2014).

26. Marlow, G. et al. Transcriptomics to study the effect of a Mediterranean-inspired diet on inflammation in Crohn's disease patients. Hum Genomics 7, 24 (2013).
27. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505, 559–563 (2014).
28. Klimenko, N. S., Odintsova, V. E., Revel-Muroz, A. & Tyakht, A. V. The hallmarks of dietary intervention-resilient gut microbiome. npj Biofilms Microbiomes 8, 77 (2022).
29. Smits, S. A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357, 802–806 (2017).
30. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. Nat Commun 8, 1162 (2017).
31. Matias, M. G., Combe, M., Barbera, C. & Mouquet, N. Ecological strategies shape the insurance potential of biodiversity. Front. Microbio. 3, (2013).
32. Hjorth, M. F. et al. Pre-treatment microbial Prevotella-to-Bacteroides ratio, determines body fat loss success during a 6-month randomized controlled diet intervention. Int J Obes 42, 580–583 (2018).
33. Luisi, M. L. E. et al. Effect of Mediterranean Diet Enriched in High Quality Extra Virgin Olive Oil on Oxidative Stress, Inflammation and Gut Microbiota in Obese and Normal Weight Adult Subjects. Front. Pharmacol. 10, 1366 (2019).
34. Millman, J. et al. Metabolically and immunologically beneficial impact of extra virgin olive and flaxseed oils on composition of gut microbiota in mice. Eur J Nutr 59, 2411–2425 (2020).

35. Ozdal, T. et al. The Reciprocal Interactions between Polyphenols and Gut Microbiota and Effects on Bioaccessibility. Nutrients 8, 78 (2016).
36. Martín-Peláez, S. et al. Effect of virgin olive oil and thyme phenolic compounds on blood lipid profile: implications of human gut microbiota. Eur J Nutr 56, 119–131 (2017).
37. Zhao, Z., Shi, A., Wang, Q. & Zhou, J. High Oleic Acid Peanut Oil and Extra Virgin Olive Oil Supplementation Attenuate Metabolic Syndrome in Rats by Modulating the Gut Microbiota. Nutrients 11, 3005 (2019).
38. Hidalgo, M. et al. Changes in Gut Microbiota Linked to a Reduction in Systolic Blood Pressure in Spontaneously Hypertensive Rats Fed an Extra Virgin Olive Oil-Enriched Diet. Plant Foods Hum Nutr 73, 1–6 (2018).
39. Chen, J. & Vitetta, L. Inflammation-Modulating Effect of Butyrate in the Prevention of Colon Cancer by Dietary Fiber. Clinical Colorectal Cancer 17, e541–e544 (2018).
40. Reichardt, N. et al. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. ISME J 8, 1323–1335 (2014).
41. Zaneveld, J. R., McMinds, R. & Vega Thurber, R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. Nat Microbiol 2, 17121 (2017).
42. Kaszubinski, S. F. et al. Dysbiosis in the Dead: Human Postmortem Microbiome Beta-Dispersion as an Indicator of Manner and Cause of Death. Front. Microbiol. 11, 555347 (2020).
43. Kumari, M. et al. Fostering next-generation probiotics in human gut by targeted dietary modulation: An emerging perspective. Food Research International 150, 110716 (2021).

44. Zhu, C. et al. Roseburia intestinalis inhibits interleukin-17 excretion and promotes regulatory T cells differentiation in colitis. Mol Med Report (2018) doi:10.3892/mmr.2018.8833.
45. Chung, W. S. F. et al. Prebiotic potential of pectin and pectic oligosaccharides to promote anti-inflammatory commensal bacteria in the human colon. FEMS Microbiology Ecology 93, (2017).
46. Martínez-González, M. A. et al. A 14-Item Mediterranean Diet Assessment Tool and Obesity Indexes among High-Risk Subjects: The PREDIMED Trial. PLoS ONE 7, e43134 (2012).
47. Scherber, R. et al. The Myeloproliferative Neoplasm Symptom Assessment Form (MPN-SAF): International Prospective Validation and Reliability Trial in 402 patients. Blood 118, 401–408 (2011).
48. Weihe, C. & Avelar-Barragan, J. Next generation shotgun library preparation for Illumina sequencing - low volume v1. doi:10.17504/protocols.io.bvv8n69w.
49. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).
50. Cantu, V. A., Sadural, J. & Edwards, R. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. https://peerj.com/preprints/27553v1 (2019) doi:10.7287/peerj.preprints.27553v1.
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).
52. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife 10, e65088 (2021).

53. Li, D. et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102, 3–11 (2016).
54. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010).
55. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47, D309–D314 (2019).
56. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol 16, 51 (2015).

TABLES AND FIGURES

Subject	Diet	MPN	Mutation	Age	Sex
2	USDA	PV	JAK2	71	M
3	USDA	MF	MPL	63	F
5	USDA	MF	JAK2	63	F
7	USDA	PV	JAK2	44	F
9	USDA	ET	JAK2	57	M
10	USDA	PV	JAK2	21	F
12	USDA	PV	JAK2	77	M
14	MED	PV	JAK2	34	F
15	MED	PV	JAK2	68	F
16	MED	ET	JAK2	70	F
17	MED	PV	JAK2	58	F
18	MED	PV	JAK2	66	M
19	USDA	ET	JAK2	61	F
20	MED	ET	CALR	71	M
21	MED	MF	CALR	25	F
22	MED	MF	JAK2	71	F
23	MED	PV	JAK2	54	F
24	MED	MF	JAK2	67	F
25	MED	PV	JAK2	59	F
26	MED	MF	JAK2	53	M
28	MED	PV	JAK2	40	M
29	MED	ET	JAK2	70	F
30	MED	PV	JAK2	50	M
31	USDA	ET	JAK2	66	F
32	USDA	PV	JAK2	67	F
33	USDA	PV	JAK2	57	F
34	USDA	MF	JAK2	51	F
35	USDA	MF	JAK2	58	F

Table 2.1: Subject characteristics. A table detailing each subject's assigned diet, MPN subtype, mutation, age, and sex.

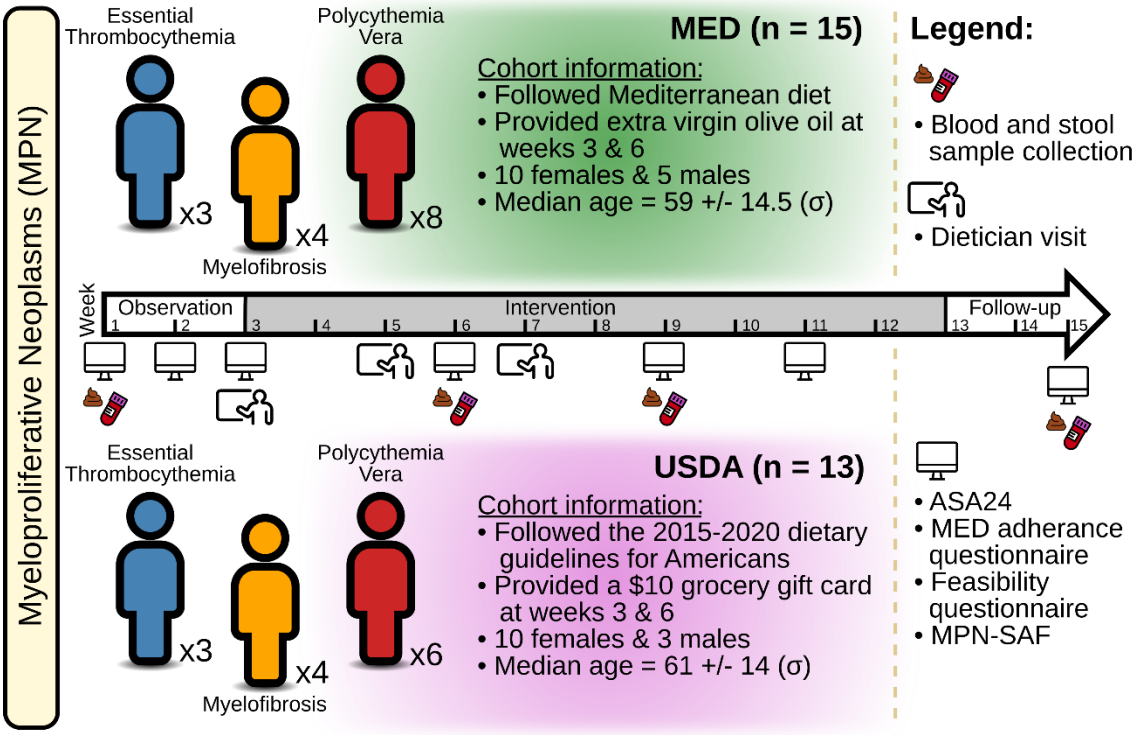


Figure 2.1: Study design. A total of 28 individuals with myeloproliferative neoplasms (MPN) were enrolled in the study. Participants were randomly assigned to dietary counseling following either a Mediterranean diet (MED, n = 15) or a conventional American diet (USDA, n = 13). The study was 15 weeks long, and had a 2-week observation period, a 10-week intervention period, and a 3-week follow-up period. Blood and stool samples were collected at weeks 1, 6, 9, and 15. At weeks 3, 5, and 7, participants met with a dietician and were informed about the core components of each diet and how to follow it. On weeks 1, 2, 3, 6, 9, 11, and 15, subjects were asked for fill out 24-hour dietary recalls (ASA24), MED adherence and feasibility questionnaires, and an MPN symptom burden assessment (MPN-SAF).

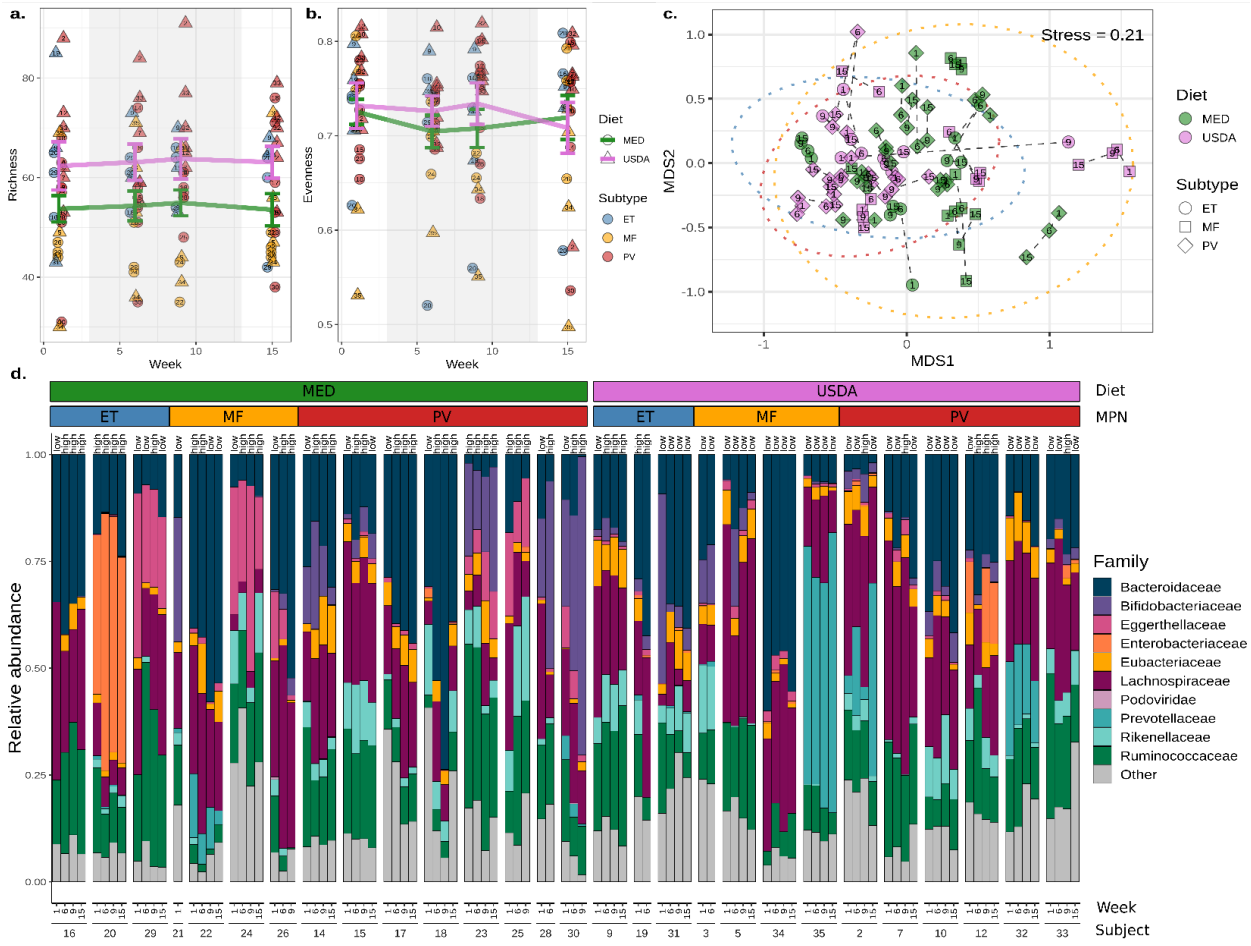


***Figure 2.2: Gut microbiome diversity and composition is stable during Mediterranean diet intervention*. A)** Microbial richness and **B)** evenness estimates of fecal samples collected at weeks 1, 6, 9, and 15. The shaded background indicates the active dietary intervention period for both diet groups. The mean richness or evenness for each group is represented with a colored line, with the error bars reflecting the standard error. Each point is labeled centrally with the individual of origin. **C)** Non-metric multidimensional scaling of Bray-Curtis dissimilarities produced from compositional microbiome data. Points are colored by diet and shaped by MPN subtype. A 95% confidence interval was drawn around each MPN subtype (Blue = ET, Yellow = MF, and Red = PV). Dashed lines connect samples taken from the same individual, and the week of collection is labeled centrally

within each point. **D)** A taxa bar plot of the top ten most abundant microbial families across individuals, time, diet, and MPN subtypes.

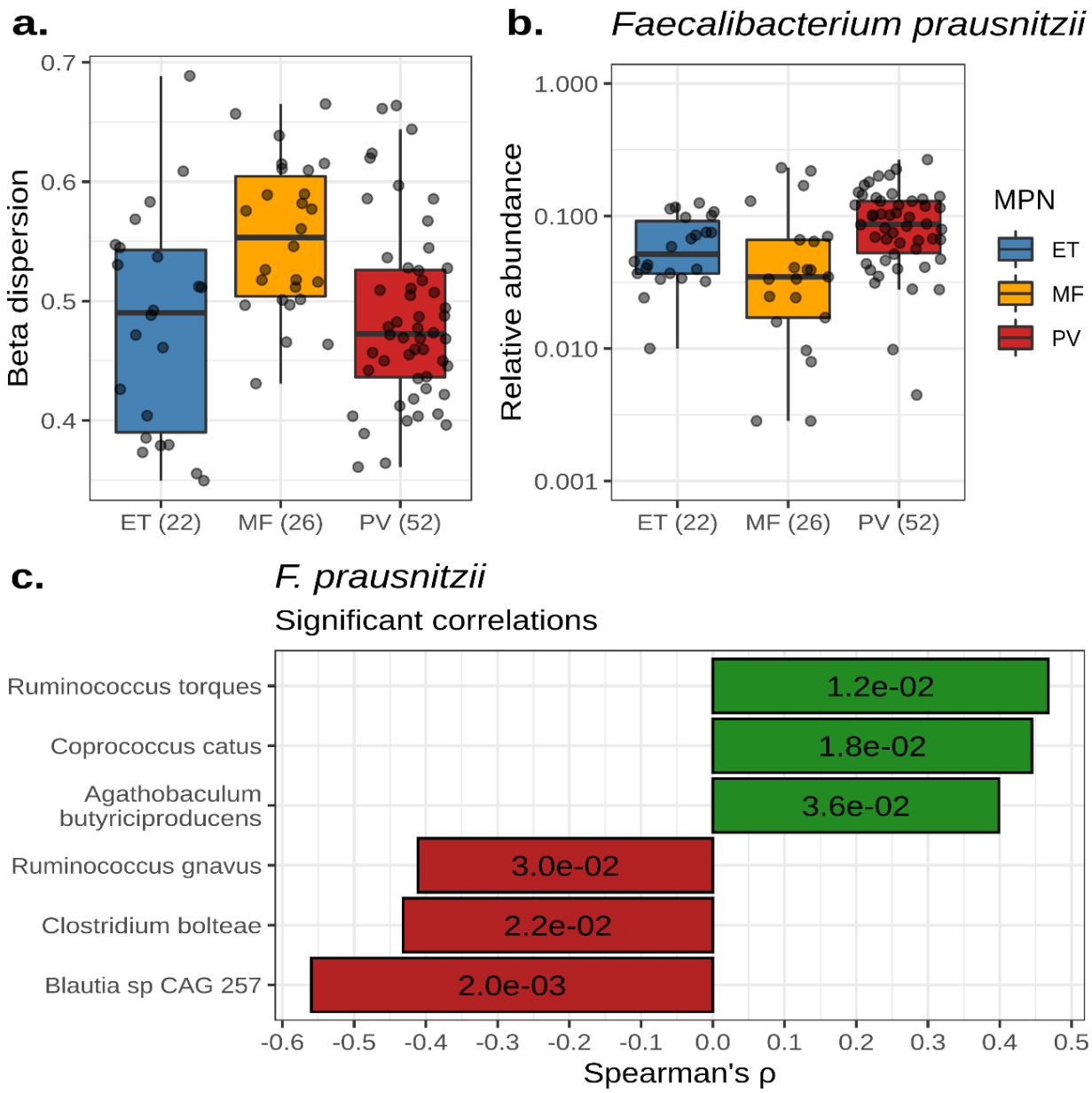


Figure 2.3: Individuals with myelofibrosis have reduced microbial diversity and altered composition. **A)** A box plot showing the beta-dispersion of each MPN subtype calculated from taxonomic Bray-Curtis dissimilarities. **B)** The relative abundance of *Faecalibacterium prausnitzii* across MPN subtypes. For **A)** and **B)** the number of samples per subtype is labeled parenthetically and the center line within each box defines the median. Boxes define the upper and lower quartiles and whiskers define 1.5x the interquartile range. **C)** A

bar plot showing the spearman correlation coefficients of microbes significantly correlated with *F. prausnitzii* abundance. P-values for each correlation are labeled within each bar.

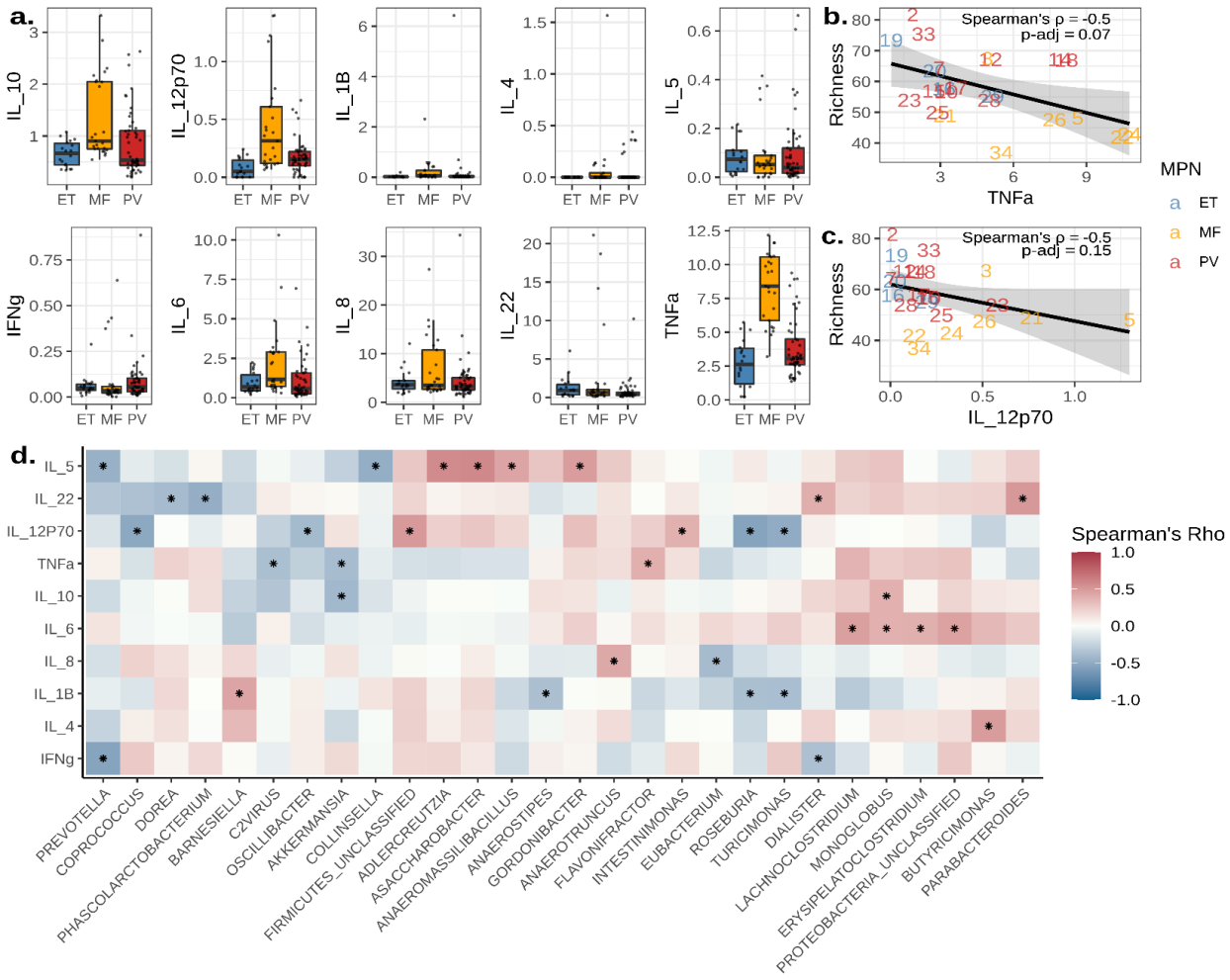
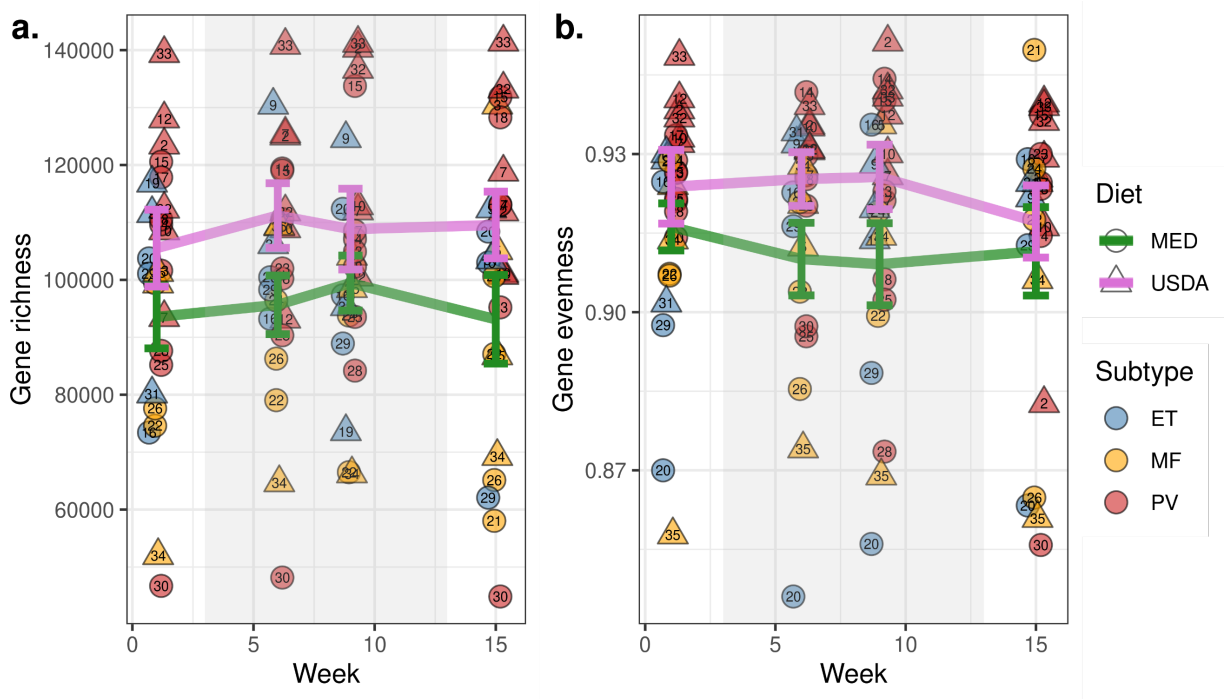


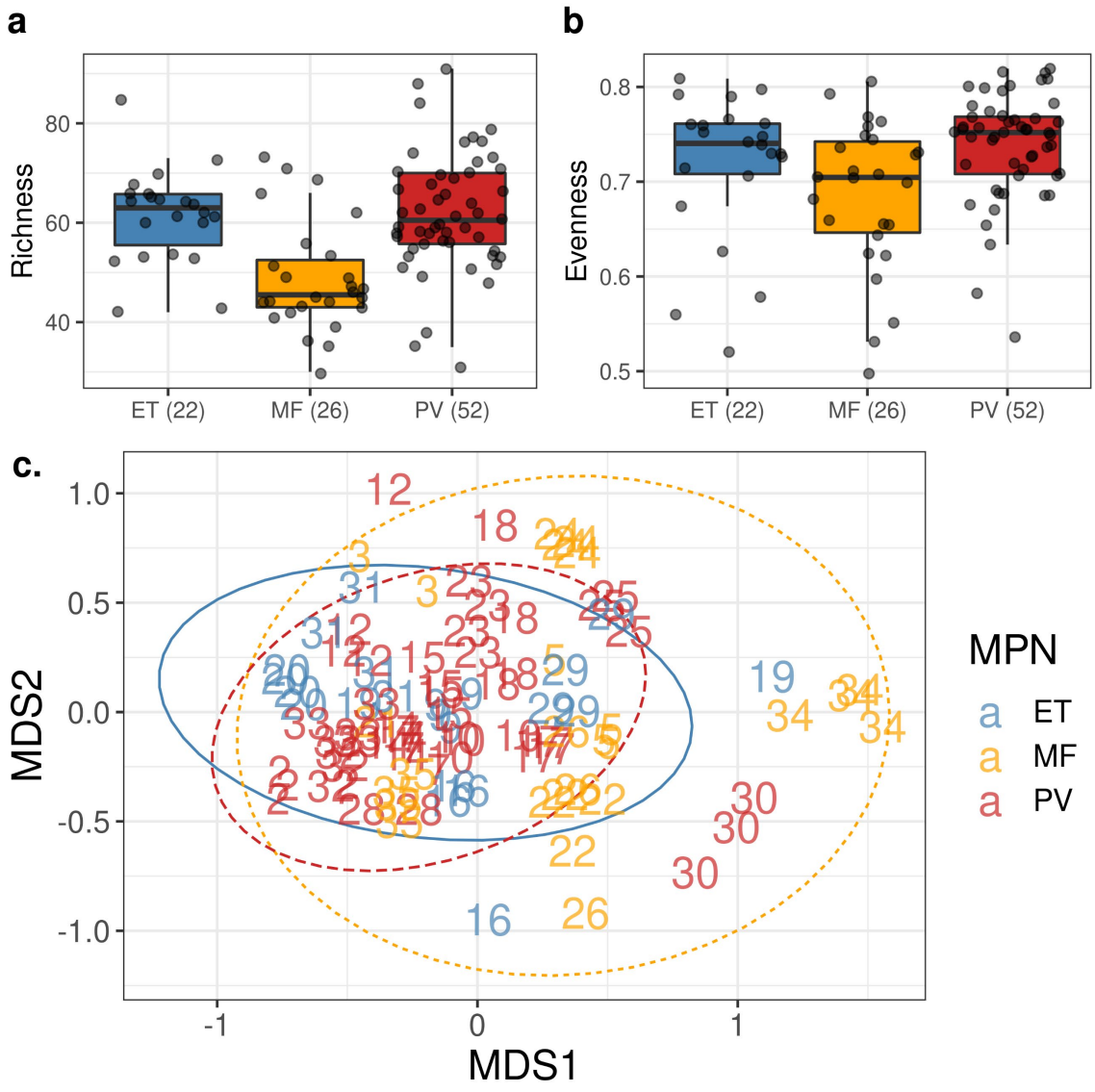
Figure 2.4: Cytokine levels are correlated with microbiome diversity and composition.

A) Box plots displaying the concentration of cytokines measured in pg/mL across MPN subtypes. The center line within each box defines the median, boxes define the upper and lower quartiles, and whiskers define 1.5x the interquartile range. **B-C)** Scatter plots of TNFα (**B**) and IL-12p70 (**C**) concentrations in pg/mL correlated with species richness estimates. Points are labeled by the individual of origin and colored by MPN subtypes. A line represents the mean, and the shaded area delineates the 95% confidence interval. **D)** A heat map of microbial genera significantly correlated with cytokine concentrations. Asterisks denote significant correlations ($p < 0.05$).

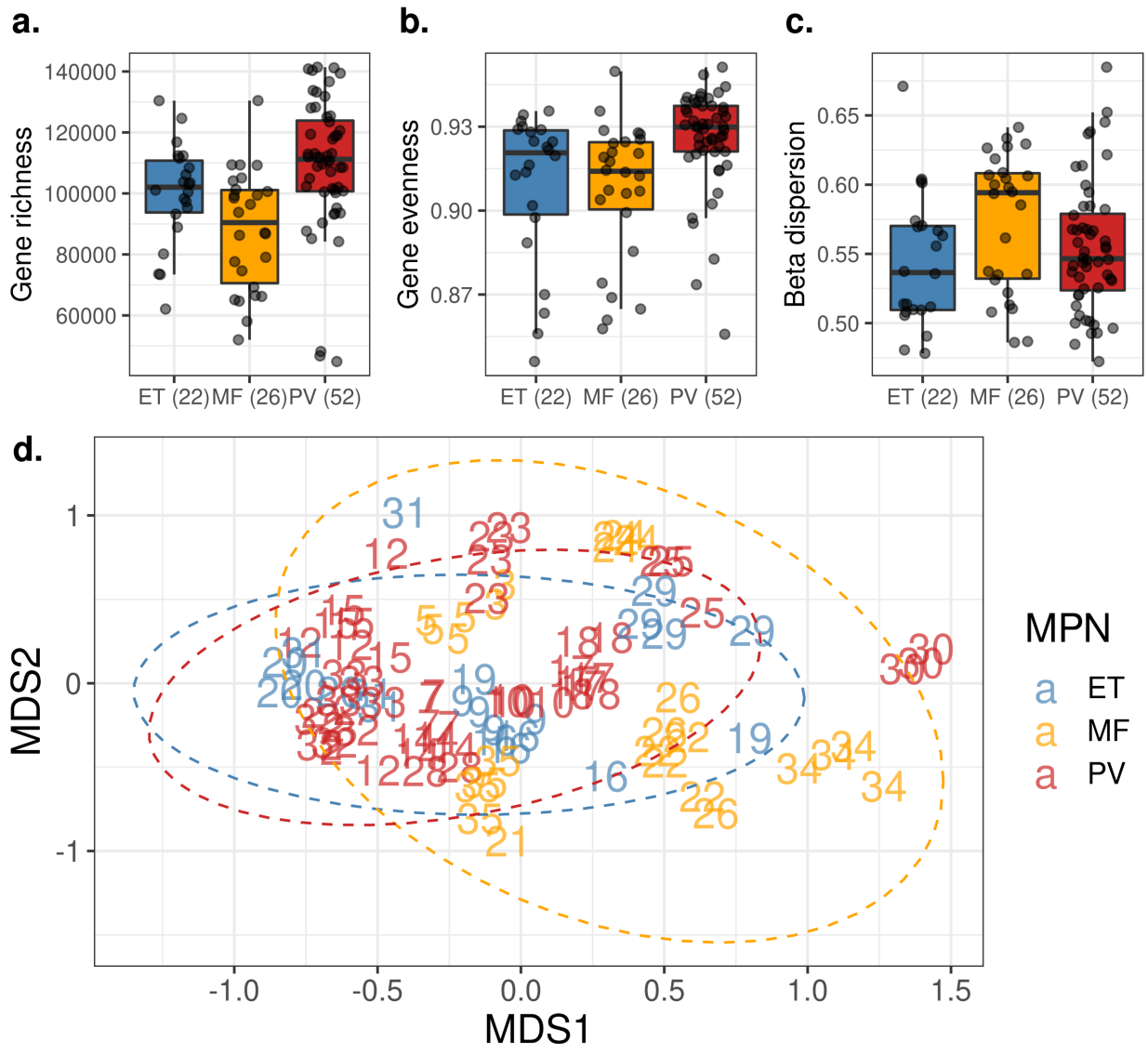
Supplementary Figures:



Supplementary Figure 2.1: **A)** Gene richness and **B)** evenness estimates of fecal samples collected at weeks 1, 6, 9, and 15. The shaded background indicates the active dietary intervention period for both diet groups. The mean richness or evenness for each group is represented with a colored line, with the error bars reflecting the standard error. Each point is labeled centrally with the individual of origin.

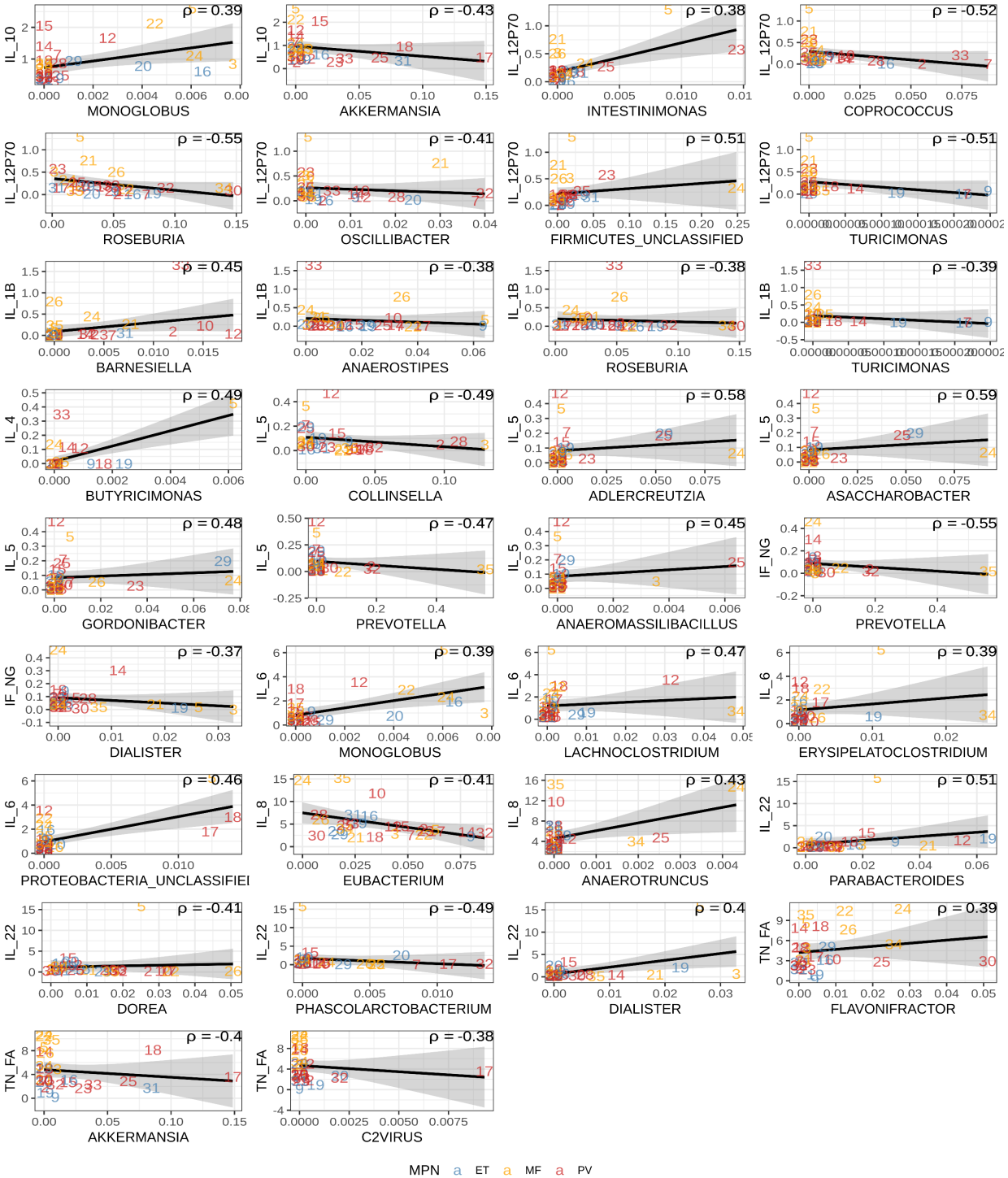


Supplementary Figure 2.2: **A** and **B)** Box plots showing microbial richness and evenness estimates across MPN subtypes. The number of samples per subtype is labeled parenthetically and the center line within each box defines the median. Boxes define the upper and lower quartiles and whiskers define 1.5x the interquartile range. **C)** Non-metric multidimensional scaling of Bray-Curtis dissimilarities produced from compositional microbiome data. Points are labeled by individual and colored by MPN subtype. A 95% confidence interval was drawn around each MPN subtype.



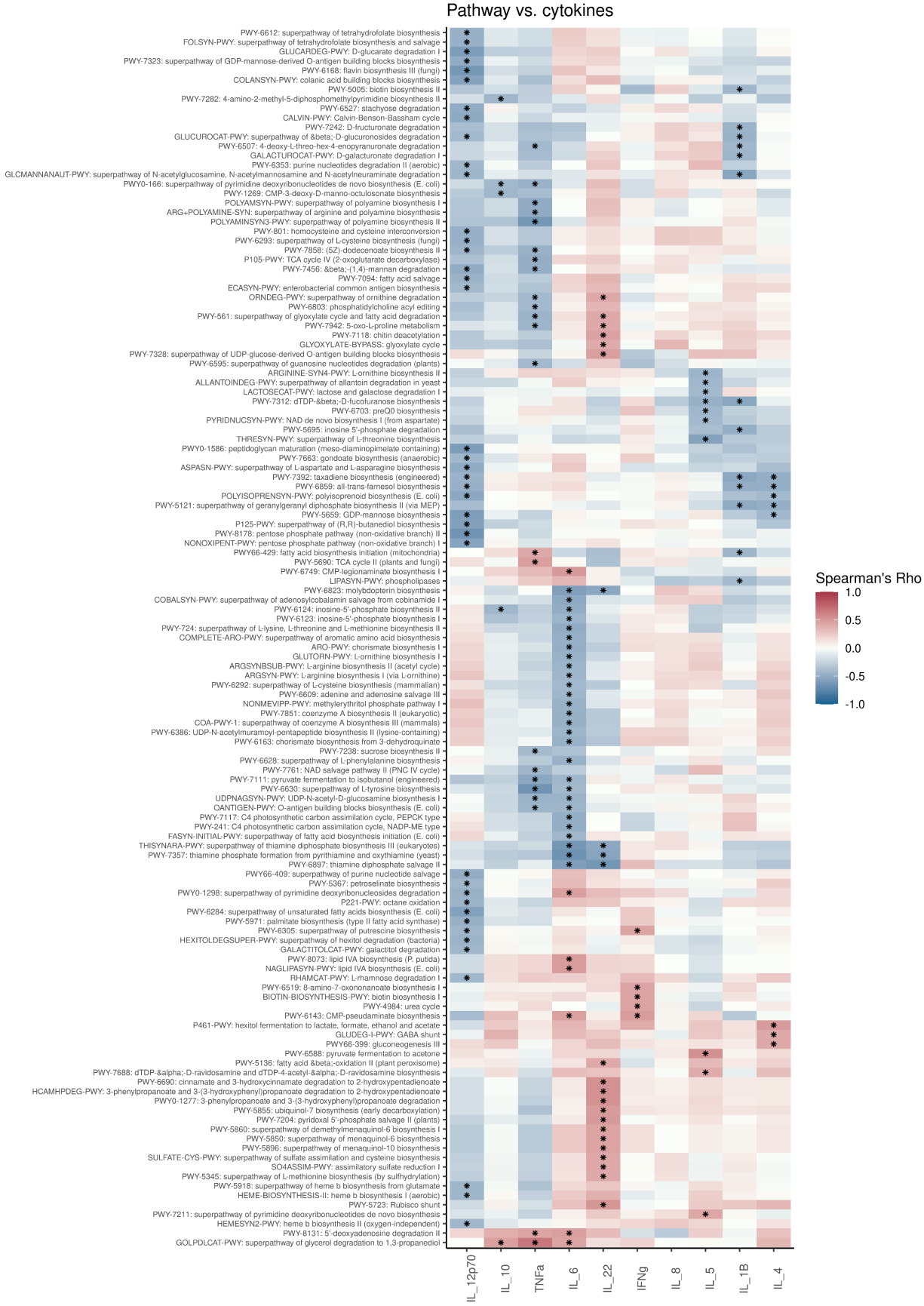
Supplementary Figure 2.3: **A** and **B)** Richness and evenness box plots of microbial genes. **C)** A box plot showing the beta dispersion of each MPN subtype calculated from gene Bray-Curtis dissimilarities. For A-C the number of samples per MPN subtype is labeled parenthetically and the center line within each box defines the median. Boxes define the upper and lower quartiles and whiskers define 1.5x the interquartile range. **D)** Non-metric multidimensional scaling of Bray-Curtis dissimilarities produced from compositional gene

data. Points are labeled by individual and colored by MPN subtype. A 95% confidence interval was drawn around each MPN subtype.

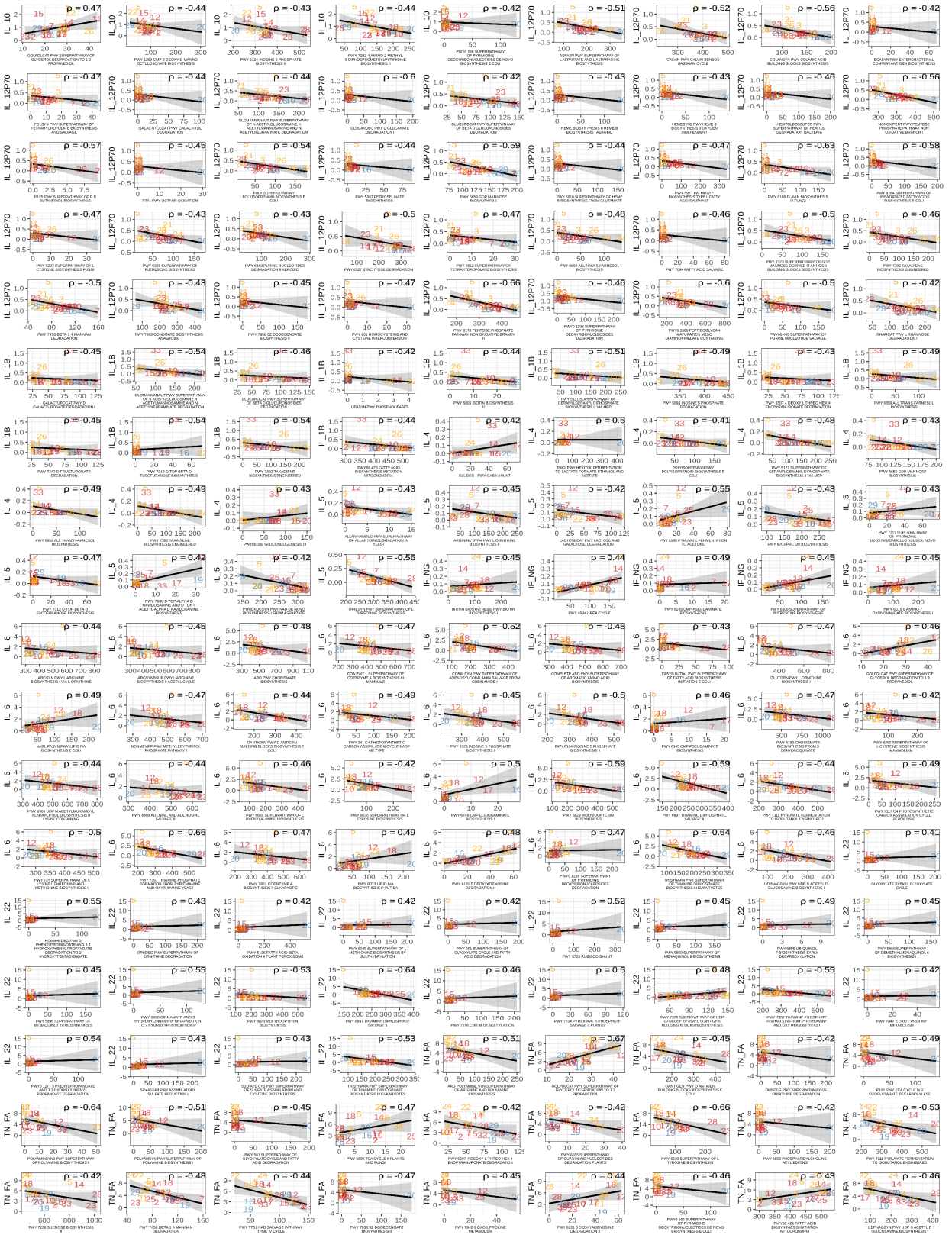


Supplementary Figure 2.4: Scatter plots of spearman correlations between the relative abundance of microbial genera and cytokine concentrations (pg/mL). Only significant correlations are shown (p-value < 0.05). The correlation coefficient is embedded in the

right of each graph. Points are labeled by the individual of origin and colored by MPN subtypes. A line represents the mean, and the shaded area delineates the 95% confidence interval.


Pathway vs. cytokines
PWY-6612: superpathway of tetrahydrofolate biosynthesis
FOLSYN-PWY: superpathway of tetrahydrofolate biosynthesis and salvage
GLUCARDEG-PWY: D-glucarate degradation I
PWY-7323: superpathway of GDP-mannose-derived O-antigen building blocks biosynthesis
PWY-6168: flavin biosynthesis III (fungi)
COLANSYN-PWY: colanic acid building blocks biosynthesis
PWY-5005: biotin biosynthesis II
PWY-7282: 4-amino-2-methyl-5-diphosphomethylpyrimidine biosynthesis II
PWY-6527: stachyose degradation
CALVIN-PWY: Calvin-Benson-Bassham cycle
PWY-7242: D-fructuronate degradation
GLUCUROCAT-PWY: superpathway of β-D-glucuronosides degradation
PWY-6507: 4-deoxy-L-threo-hex-4-enopyranuronate degradation
GALACTUROCAT-PWY: D-galacturonate degradation I
PWY-6353: purine nucleotides degradation II (aerobic)
GLCMANNANAUT-PWY: superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation
PWY0-166: superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis (E. coli)
PWY-1269: CMP-3-deoxy-D-manno-octulosonate biosynthesis
POLYAMSYN-PWY: superpathway of polyamine biosynthesis I
ARG+POLYAMINE-SYN: superpathway of arginine and polyamine biosynthesis
POLYAMINSYN3-PWY: superpathway of polyamine biosynthesis II
PWY-801: homocysteine and cysteine interconversion
PWY-6293: superpathway of L-cysteine biosynthesis (fungi)
PWY-7858: (5Z)-dodecenoate biosynthesis II
P105-PWY: TCA cycle IV (2-oxoglutarate decarboxylase)
PWY-7456: β-(1,4)-mannan degradation
PWY-7094: fatty acid salvage
ECASYN-PWY: enterobacterial common antigen biosynthesis
ORNDEG-PWY: superpathway of ornithine degradation
PWY-6803: phosphatidylcholine acyl editing
PWY-561: superpathway of glyoxylate cycle and fatty acid degradation
PWY-7942: 5-oxo-L-proline metabolism
PWY-7118: chitin deacetylation
GLYOXYLATE-BYPASS: glyoxylate cycle
PWY-7328: superpathway of UDP-glucose-derived O-antigen building blocks biosynthesis
PWY-6595: superpathway of guanosine nucleotides degradation (plants)
ARGININE-SYN4-PWY: L-ornithine biosynthesis II
ALLANTOINDEG-PWY: superpathway of allantoin degradation in yeast
LACTOSECAT-PWY: lactose and galactose degradation I
PWY-7312: dTDP-β-D-fucofuranose biosynthesis
PWY-6703: preQ0 biosynthesis
PYRIDNUCSYN-PWY: NAD de novo biosynthesis I (from aspartate)
PWY-5695: inosine 5'-phosphate degradation
THRESYN-PWY: superpathway of L-threonine biosynthesis
PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)
PWY-7663: gondoate biosynthesis (anaerobic)
ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis
PWY-7392: taxadiene biosynthesis (engineered)
PWY-6859: all-trans-farnesol biosynthesis
POLYISOPRENSYN-PWY: polyisoprenoid biosynthesis (E. coli)
PWY-5121: superpathway of geranylgeranyl diphosphate biosynthesis II (via MEP)
PWY-5659: GDP-mannose biosynthesis
P125-PWY: superpathway of (R,R)-butanediol biosynthesis
PWY-8178: pentose phosphate pathway (non-oxidative branch) II
NONOXIPENT-PWY: pentose phosphate pathway (non-oxidative branch) I
PWY66-429: fatty acid biosynthesis initiation (mitochondria)
PWY-5690: TCA cycle II (plants and fungi)
PWY-6749: CMP-legionaminate biosynthesis I
LIPASYN-PWY: phospholipases
PWY-6823: molybdopterin biosynthesis
COBALSYN-PWY: superpathway of adenosylcobalamin salvage from cobinamide I
PWY-6124: inosine-5'-phosphate biosynthesis II
PWY-6123: inosine-5'-phosphate biosynthesis I
PWY-724: superpathway of L-lysine, L-threonine and L-methionine biosynthesis II
COMPLETE-ARO-PWY: superpathway of aromatic amino acid biosynthesis
ARO-PWY: chorismate biosynthesis I
GLUTORN-PWY: L-ornithine biosynthesis I
ARGSYNBSUB-PWY: L-arginine biosynthesis II (acetyl cycle)
ARGSYN-PWY: L-arginine biosynthesis I (via L-ornithine)
PWY-6292: superpathway of L-cysteine biosynthesis (mammalian)
PWY-6609: adenine and adenosine salvage III
NONMEVIPP-PWY: methylerythritol phosphate pathway I
PWY-7851: coenzyme A biosynthesis II (eukaryotic)
COA-PWY-1: superpathway of coenzyme A biosynthesis III (mammals)
PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)
PWY-6163: chorismate biosynthesis from 3-dehydroquinate
PWY-7238: sucrose biosynthesis II
PWY-6628: superpathway of L-phenylalanine biosynthesis
PWY-7761: NAD salvage pathway II (PNC IV cycle)
PWY-7111: pyruvate fermentation to isobutanol (engineered)
PWY-6630: superpathway of L-tyrosine biosynthesis
UDPNAGSYN-PWY: UDP-N-acetyl-D-glucosamine biosynthesis I
OANTIGEN-PWY: O-antigen building blocks biosynthesis (E. coli)
PWY-7117: C4 photosynthetic carbon assimilation cycle, PEPCK type
PWY-241: C4 photosynthetic carbon assimilation cycle, NADP-ME type
FASYN-INITIAL-PWY: superpathway of fatty acid biosynthesis initiation (E. coli)
THISYNARA-PWY: superpathway of thiamine diphosphate biosynthesis III (eukaryotes)
PWY-7357: thiamine phosphate formation from pyrithiamine and oxythiamine (yeast)
PWY-6897: thiamine diphosphate salvage II
PWY66-409: superpathway of purine nucleotide salvage
PWY-5367: petroselinate biosynthesis
PWY0-1298: superpathway of pyrimidine deoxyribonucleosides degradation
P221-PWY: octane oxidation
PWY-6284: superpathway of unsaturated fatty acids biosynthesis (E. coli)
PWY-5971: palmitate biosynthesis (type II fatty acid synthase)
PWY-6305: superpathway of putrescine biosynthesis
HEXITOLDEGSUPER-PWY: superpathway of hexitol degradation (bacteria)
GALACTITOLCAT-PWY: galactitol degradation
PWY-8073: lipid IVA biosynthesis (P. putida)
NAGLIPASYN-PWY: lipid IVA biosynthesis (E. coli)
RHAMCAT-PWY: L-rhamnose degradation I
PWY-6519: 8-amino-7-oxononanoate biosynthesis I
BIOTIN-BIOSYNTHESIS-PWY: biotin biosynthesis I
PWY-4984: urea cycle
PWY-6143: CMP-pseudaminate biosynthesis
P461-PWY: hexitol fermentation to lactate, formate, ethanol and acetate
GLUDEG-I-PWY: GABA shunt
PWY66-399: gluconeogenesis III
PWY-6588: pyruvate fermentation to acetone
PWY-5136: fatty acid β-oxidation II (plant peroxisome)
PWY-7688: dTDP-α-D-ravidosamine and dTDP-4-acetyl-α-D-ravidosamine biosynthesis
PWY-6690: cinnamate and 3-hydroxycinnamate degradation to 2-hydroxypentadienoate
HCAMHPDEG-PWY: 3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate degradation to 2-hydroxypentadienoate
PWY0-1277: 3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate degradation
PWY-5855: ubiquinol-7 biosynthesis (early decarboxylation)
PWY-7204: pyridoxal 5'-phosphate salvage II (plants)
PWY-5860: superpathway of demethylmenaquinol-6 biosynthesis I
PWY-5850: superpathway of menaquinol-6 biosynthesis
PWY-5896: superpathway of menaquinol-10 biosynthesis
SULFATE-CYS-PWY: superpathway of sulfate assimilation and cysteine biosynthesis
SO4ASSIM-PWY: assimilatory sulfate reduction I
PWY-5345: superpathway of L-methionine biosynthesis (by sulfhydrylation)
PWY-5918: superpathway of heme b biosynthesis from glutamate
HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)
PWY-5723: Rubisco shunt
PWY-7211: superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis
HEMESYN2-PWY: heme b biosynthesis II (oxygen-independent)
PWY-8131: 5'-deoxyadenosine degradation II
GOLPDLCAT-PWY: superpathway of glycerol degradation to 1,3-propanediol
IL_12p70
IL_10
TNFa
IL_6
IL_22
IFNg
IL_8
IL_5
IL_1B
IL_4
Spearman's Rho
1.0
0.5
0.0
-0.5
-1.0

Supplementary Figure 2.5: A heat map of metabolic pathways significantly correlated with cytokine concentrations. Asterisks denote significant correlations ($p < 0.05$).


MPN a ET a MF a PV

Supplementary Figure 2.6: Scatter plots of spearman correlations between the abundance of metabolic pathways (counts per million) and cytokine concentrations (pg/mL). Only significant correlations are shown (p-value < 0.05). The correlation coefficient is embedded in the right of each graph. Points are labeled by the individual of origin and colored by MPN subtypes. A line represents the mean, and the shaded area delineates the 95% confidence interval.

Supplementary Tables

	Df	SumOfSqs	R2	F	Pr(>F)
Diet	1	0.410168	0.057337	1.459795	0.046
Residual	24	6.743428	0.942663	NA	NA
Total	25	7.153596	1	NA	NA

PERMANOVA formula: OTU_table[week1,] ~ Diet

Supplementary Table 2.1: A table of results produced by PERMANOVA using only samples from week 1, demonstrating significant differences in the microbiome between diet groups pre-intervention.

	Df	SumOfSqs	R2	F	Pr(>F)
Week	3	0.199089	0.007196	0.231956	0.764
Residual	96	27.46576	0.992804	NA	NA
Total	99	27.66485	1	NA	NA

PERMANOVA formula: OTU_table ~ Week, strata = Subject

Supplementary Table 2.2: A table of results produced by PERMANOVA, which tested whether samples across all four time points (Weeks 1, 6, 9, and 15) significantly differed in microbial composition within each individual.

ET	Df	SumOfSqs	R2	F	Pr(>F)
Week	3	0.240761	0.045829	0.288184	0.631
Residual	18	5.012647	0.954171	NA	NA
Total	21	5.253408	1	NA	NA
MF	Df	SumOfSqs	R2	F	Pr(>F)
Week	3	0.210731	0.026236	0.197581	0.603
Residual	22	7.8214	0.973764	NA	NA
Total	25	8.032131	1	NA	NA
PV	Df	SumOfSqs	R2	F	Pr(>F)
Week	3	0.197215	0.01567	0.254711	0.769
Residual	48	12.38828	0.98433	NA	NA
Total	51	12.58549	1	NA	NA

PERMANOVA formula: OTU_table[ET/MF/PV,] ~ Week, strata = Subject

Supplementary Table 2.3: A table of results produced by PERMANOVA, which tested whether samples across all four time points (Weeks 1, 6, 9, and 15) significantly differed in microbial composition within each individual one MPN subtype at a time.

	Df	SumOfSqs	R2	F	Pr(>F)
Week	3	0.303795	0.009031	0.291641	0.793
Residual	96	33.33357	0.990969	NA	NA
Total	99	33.63736	1	NA	NA

PERMANOVA formula: Gene_table ~ Week, strata = Subject

Supplementary Table 2.4: A table of results produced by PERMANOVA, which tested whether samples across all four time points (Weeks 1, 6, 9, and 15) significantly differed in microbial gene composition within each individual.

	Df	SumOfSqs	R2	F	Pr(>F)
dna_extraction	3	4.066052	0.146975	23.30149	0.001
library_prep	1	0.789917	0.028553	13.58044	0.001
Age	1	0.929177	0.033587	15.97461	0.001
Sex	1	0.91958	0.03324	15.80963	0.001
Diet	1	0.336291	0.012156	5.781588	0.001
MPN	2	1.687195	0.060987	14.50332	0.001
MPN:Subject	20	14.86503	0.537325	12.77814	0.001
Residual	70	4.071609	0.147176	NA	NA
Total	99	27.66485	1	NA	NA

PERMANOVA formula: OTU_table ~ dna_extraction + library_prep + Age + Sex + Diet + MPN / Subject, strata = Week

Supplementary Table 2.5: A table of results produced by PERMANOVA, investigating the association of all available factors with microbial composition. Only time points from the same week were compared with each other, and subjects nested within an MPN subtype.

	Df	SumOfSqs	R2	F	Pr(>F)
dna_extraction	3	4.188348	0.124515	16.54748	0.001
library_prep	1	0.953413	0.028344	11.30034	0.001
Age	1	1.093623	0.032512	12.96218	0.001
Sex	1	1.064363	0.031642	12.61537	0.001
MPN	2	2.268921	0.067452	13.44621	0.001
MPN:Subject	21	18.16277	0.539958	10.25116	0.001
Residual	70	5.905922	0.175576	NA	NA
Total	99	33.63736	1	NA	NA

PERMANOVA formula: Gene_table ~ dna_extraction + library_prep + Age + Sex + Diet + MPN / Subject, strata = Week

Supplementary Table 2.6: A table of results produced by PERMANOVA, investigating the association of all available factors with microbial gene composition. Only time points from the same week were compared with each other, and subjects nested within an MPN subtype.

CHAPTER 3

Breaking it down: *In vitro* cultivation of human gut samples with dietary fiber

Authors: Julio Avelar-Barragan, Zachary Pashkutz, Carlos Ferrer Inguito, Roaa Saadeh, Soumaya El Halas, Katrine Whiteson

ABSTRACT

The gut microbiome has become increasingly recognized as an important modulator of human health. As such, there is a growing interest in developing strategies to manipulate the gut microbiome to promote health and prevent disease. One potential method which can be used is dietary fiber supplementation. Fiber resists host digestion and is instead fermented by gut microbes to produce beneficial metabolites, like short-chain fatty acids (SCFAs). However, not all fibers are created equal. Different fibers contain unique monosaccharides and oligosaccharide branching patterns. Therefore, only specific microbes can ferment specific fibers. Currently, it is unclear how specific fibers are utilized by the gut microbiota. Here, we investigated how three common dietary fibers, inulin, pectin, and psyllium, were fermented by the fecal microbiota of 15 healthy subjects from the U.S. and 15 healthy subjects from Morocco. Fermentation was assessed *in vitro* using fecal community cultures grown anaerobically at 37 °C. We performed shotgun metagenomic sequencing on cultures at 0h and 24h to characterize the gut microbiome in the presence and absence of fiber. Cultures became dominated by a subset of microbes over time, as illustrated by decreases in alpha-diversity from 0h to 24h. Microbial composition was significantly dependent on the country, subject of origin, and fiber treatment. Pectin and psyllium were able to significantly enrich for microbes, but not

inulin. Together, this research seeks to advance the use of dietary fiber as a tool for microbiome manipulation.

INTRODUCTION

Many gut microbes and their metabolites have been associated with health and disease; therefore, manipulation of the gut microbiome is a potential strategy for promoting health. One method to alter the gut microbiome is by performing high fiber dietary interventions, as diet is important in shaping the diversity and composition of microbes in the gut. Fiber is defined as edible carbohydrate polymers with three or more monomeric units that are resistant to the endogenous digestive enzymes. It is fermented by gut microbes to produce short-chain fatty acids (SCFAs) and other beneficial metabolites. [1] High dietary fiber consumption correlates with a reduced risk of developing type II diabetes, cardiovascular disease, obesity, and various gastrointestinal diseases, like colorectal cancer. [2] Diseases associated with low fiber consumption are more common in industrialized countries and correlates with differences in gut microbiome composition when compared to non-industrialized countries. [3]

Previous studies conducting high dietary fiber interventions have shown mixed results, with some reporting beneficial health outcomes and others reporting no changes in the gut microbiome. [4] It is not known how specific fibers are utilized by gut microbes, and how their fermentation produces health benefits. There are several characteristics which have been postulated to affect the gut microbiome's ability to utilize dietary fiber. One is the physiochemical complexity of fiber. [5] The enzymes which are used to digest fiber are specific to the type of polysaccharides, linkage types, and branching patterns. For example,

type II rhamnogalacturonans, which are found in pectin, are complex fibers with 21 types of distinct glycosidic linkages. [6] Ndeh, *et al.* demonstrated that of 29 *Bacteroidetes* species tested, only *Bacteroides thetaiotaomicron* utilized this fiber. [7] On the other hand, fructooligosaccharides, which are structurally simple, can be processed by many *Firmicutes*, *Bifidobacterium*, and *Bacteroides* species. [8] Although these studies support the hypothesis that more complex fibers are utilized by fewer microbes; more studies are needed to test a greater diversity of fibers and community compositions.

Other factors which affect a microbiome's ability to ferment fiber is the frequency of exposure to fiber and the diversity of microbes present. [6] If a specific fiber is rarely consumed, it is likely that few microbes will be able to digest it, as it is not evolutionarily beneficial to maintain genes which confer a fitness cost but provide little benefit. To date, dietary fiber intervention studies have provided some insight into the dynamics between the microbiota and fiber, but more studies are needed to isolate the effects of fiber without host influence.

To further understand how the gut microbiota utilizes fiber, we cultured the fecal microbial communities of 15 healthy individuals from the U.S. and 15 Moroccan individuals supplemented with and without various dietary fibers. Fibers used included inulin, pectin, and psyllium husks due to their ubiquity in food and varying degrees of resistance to microbial fermentation. Inulin is a fructose polymer held together by β-(2,1) linkages and it present in common foods like bananas, onion, wheat, and artichokes. [9] Pectin consists of complex homogalacturonans and type I and II rhamnogalacturonans, and it is common in citrus peels, apples, and apricots. [6] Psyllium husk is rich in arabinoxylan, which is a

hemicellulose consisting of β-(1,4) xylose residues with arabinose substitutions and is present in fiber supplements. [10] With this research, we aim to answer the following questions: 1) Can fiber be used to selectively enrich for microbes? 2) If so, is the direction of taxonomic or carbohydrate active enzyme enrichment specific to each fiber and its structural complexity? 3) Lastly, does the variation between cohorts impact the microbiome's response to fiber? An important future direction we aim to include when we publish this work are measures of pH and SCFAs.

RESULTS

Study design

For this study, we recruited 30 individuals, 15 from the U.S. and 15 from Morocco. Each provided a fecal sample which was then used for culturing. There were four treatments, including cultures without fiber and those with either inulin, pectin, or psyllium. To mimic the transit time of fiber in the large intestine, we allowed our cultures to ferment for 24 hours anaerobically at 37 °C in BHI media. Samples were collected before culturing and after 24 hours. Afterwards, samples were subjected to shotgun metagenomic sequencing and gas-chromatography mass spectrometry (Figure 3.1). Sequences were assembled and microbes were defined as metagenome assembled genomes (MAG).

Country, time, and fiber treatment are all significantly associated with microbial and enzymatic diversity and composition

We hypothesized that the culturing would select for microbes which are able to grow rapidly using the given carbon source. After 24 hours of fermentation, assessment of microbial taxonomy suggested that U.S. samples often became dominated by

Bifidobacterium, while Moroccan samples often became dominated with *Clostridium* and *Paraclostridium* (Figure 3.2). This was illustrated by a decrease in microbial and carbohydrate active enzyme (CAZyme) Shannon diversities at 24 hours (Figure 3.3A and 3D). Among culture-free samples, linear mixed effects (LME) testing determined that the U.S. had a significantly higher microbial diversity than Morocco (LME p-value = 0.04). Within U.S. samples, inulin and pectin treated samples had decreased Shannon diversities compared to controls (Inulin LME p-value = 0.002, Pectin LME p-value = 0.055). In Moroccan samples, pectin treated ones had a marginally higher Shannon diversity when compared to controls (LME p-value = 0.071). CAZyme Shannon diversity was not significantly different between cohorts at time point 0. Within cohorts, only inulin treated U.S. samples had significantly higher CAZyme diversity compared to controls (LME p-value = 0.012). We also calculated the average microbial genome size per sample and discovered U.S. samples had significantly larger genomes when compared to Moroccan samples at 0h (LME p-value < 2.22×10^{-16}, Supplementary Figure 3.1).

Analysis of microbial composition was performed using non-metric multidimensional scaling (NMDS) and permutational multivariate analysis of variance (PERMANOVA). At T0, NMDS displayed separate clustering by cohort, which was significantly associated with ~15% of the variance associated with the microbiome (PERMANOVA p-value < 0.001, Figure 3.3B). The subject of origin was associated with 67% of the microbiome-associated variance at T0 (PERMANOVA p-value < 0.001). The individual and cohort variation of samples decreased to 46% and 11% at T24, respectively (PERMANOVA p-value < 0.001 for both). No immediate clustering based on fiber treatment was observed within the Moroccan cohort. U.S. samples were more disparate at 24 hours

(Figure 3.3C). Fiber treatment was significantly associated with 21% of the microbiome variance in both cohorts (PERMANOVA p-value < 0.001 for U.S. and PERMANOVA p-value < 0.023 for Moroccan).

We next examined CAZyme composition at T0 and found separation of samples by country using NMDS (Figure 3.3E). After 24 hours, samples remained loosely clustered by country (Figure 3.3F). PERMANOVA demonstrated that country of origin was significantly associated with 5.7% of the CAZyme variance at T0 (PERMANOVA p-value < 0.001). Most of the variation at T0 was due to the individual, with an R^2 of 57% (PERMANOVA p-value < 0.001). After fermenting for 24 hours, the country of origin remained at 5%, while the individual decreased to 22%. Meanwhile, fiber treatment explained 35% of observed variance in CAZyme composition at T24 (PERMANOVA p-value < 0.001 for all).

Pectin and psyllium enrich for microbes, but inulin does not

At time point 0, 131 microbes were significantly different between U.S. and Moroccan samples (Figure 3.4A). Of these, 68 were more abundant in the U.S., and 63 were more abundant in Morocco. The most significantly different microbe in the U.S. was *Alistipes putredinis* (q-value = 6.22×10^{-7}) and *Ligilactobacillus ruminis* (q-value = 1.56×10^{-7}) in Moroccan samples. Of the 68 microbes more abundant in the U.S., 33 of these were *Lachnospiraceae* (Figure 3.4B). The 63 microbes more abundant in the Moroccan cohort were represented by *Lachnospiraceae, Bifidobacteriaceae*, *Anaerovoracaceae,* and others.

Differential abundance analysis across all samples at 24 hours revealed 78 microbes that were significantly enriched in fiber treated samples when compared to controls (Figure 3.4C). Of those, 69 were enriched with pectin, and 9 were enriched with psyllium

husk. Stratification of samples by cohort, followed by differential abundance analysis, showed that 22 microbes were enriched by fiber in U.S. samples and 80 microbes were enriched by fiber in the Moroccan ones. Within the U.S. cohort, 12 of the significant different microbes were enriched by pectin, and 10 were enriched by psyllium husk. Within the Moroccan cohort, 79 of the significant different microbes were enriched by pectin, and 1 was enriched by psyllium husk. The largest taxonomic group enriched by both fibers in both cohorts consisted of the *Lachnospiraceae* and *Oscillospiraceae.* The exception was pectin supplemented U.S. samples, whose most frequently enriched taxa was the *Bifidobacteriaceae* (Figure 3.4D). No microbes were enriched by inulin in both cohorts.

Characterization of microbial Carbohydrate Active Enzymes

Next, we wanted to investigate the CAZymes within microbes. The microbe encoding the greatest number of unique CAZymes was *Bacteroides intestinalis,* with 335 unique CAZymes. Within the top ten, six of the microbes encoding the most unique CAZymes belonged to the genus *Bacteroides* (Supplementary Figure 3.2). The most common CAZyme families were glycoside hydrolases (GH), followed by glycosyl transferases (GT). Across specific CAZyme subfamilies, GT2, GH2, GT4, GH3, and GT51 were the most frequently encoded, in that order (Figure 3.5A). Within microbes which were significantly enriched by pectin, GT2, GH3, GH2, GT35, and GT51 were the most common CAZymes, while GT2, GH2, GH3, GT4, and GT51 were the most common in psyllium-enriched microbes (Figure 3.5B and 5C). Differential abundance testing of CAZymes did not resolve any significantly enriched features after correcting for multiple hypothesis testing.

Since multiple *Bifidobacterium* and *Blautia* species were significantly enriched by fiber, we next explored their pangenomes to potentially identify CAZymes implicated in pectin or psyllium utilization. In the *Bifidobacterium* core genome, 5 CAZymes were found (Figure 3.6). The remaining CAZymes were found on the species-specific regions of the genome. *Bifidobacterium animalis* had an additional 11 CAZymes. Conversely, no CAZymes were found in the species-specific regions of the *B. angulatum, B. bifidum,* and *B. catenulatum* genomes. Within the *Blautia* core genome, we detected 14 CAZymes (Figure 3.7). *Blautia faecis* had an additional 7 CAZymes in its species-specific region of the genome, while *B. luti, B. sp000285855, B. sp003477525, B. sp000436615,* and *B. sp900548245* had no additional CAZymes.

DISCUSSION

In this study, we used fecal samples from 15 U.S. and 15 Moroccan individuals for *in vitro* microbial community culturing, with and without dietary fiber, to investigate: 1) If fiber could be used to selectively enrich for microbes, 2) If the direction of taxonomic or carbohydrate active enzyme enrichment was specific to each fiber and its structural complexity, and 3) If the variation between cohorts impacted the microbiome's response to fiber. We discovered significant differences in the initial microbiome diversity and composition of both cohorts. After 24-hours, cohorts continued to diverge in microbial composition both in the presence and absence of various fibers. The addition of fiber had a significant impact on microbial diversity and composition after fermentation. Specifically, pectin supplementation resulted in the greatest number of significantly enriched microbes overall when compared to no fiber controls. Psyllium husk supplementation resulted in the

second greatest number of enriched microbes, while inulin supplementation provided no enrichment. Moroccan microbes were more frequently enriched by pectin, but the most significant enrichments occurred with microbes found in U.S. samples.

Our data supports the hypothesis that fibers can be used to selectively enrich for microbes. Importantly, the number of significantly different microbes produced by pectin, psyllium, and inulin correlated with fiber structural complexity. It has been hypothesized that more structurally complex fibers enrich for more specific microorganisms because they need to encode all the required CAZymes for fermentation. [5] Pectin is one of the most complex plant polysaccharides, with three main types, and it produced the most enrichments with 69 microbial taxa. The first and most common pectin fibers are homogalacturonans, which are polymers of α (1,4)-linked d-galacturonic acids. [6] Next are type-I rhamnogalacturonans, which have an alpha linked d-galacturonic acid and l-rhamnose backbone with arabinan, galactan, or arabinogalactan sidechains. [6] Lastly, there are type-II rhamnogalacturonans, which have a homogalacturonan backbone and sidechains with 13 different types of sugars and 21 different linkages. [6] Psyllium, by comparison, enriched 9 microbes. Psyllium is primarily made of arabinoxylan, which is a highly branched fiber containing both β (1,4) and β (1,3) glycosidic linkages in its xylan backbone. [10] Inulin is the least complex fiber in this study, consisting of mainly fructose monomers joined by β (2,1) linkages. [9] Inulin did not produce any significantly enriched microbes.

Prior studies performing pectin enrichments have seen increases in the abundance of *Bacteroides, Firmicutes, Bifidobacterium, Lachnospiraceae,* and *Oscillospiraceae.* [7, 14-16] To

date, the only known pectin degrading microorganisms belong largely to the *Bacteroides* and *Firmicutes* species, as only they encode the complete repertoire of necessary CAZymes. [6] These CAZymes include GH28, GH78, GH105, GH106, CE8, CE12, PL1, PL9, PL10, PL11, and PL22. [6] *Bifidobacteria* are unable to grow in media with pectin as the sole carbon source, and thus, their enrichment is likely caused by the fermentation of downstream metabolites produced by pectolytic microorganisms. [6] Unlike pectin, psyllium produced fewer significant enrichments and mostly within U.S. samples. All the enriched microbes were *Lachnospiraceae*. Other studies investigating the effect of psyllium supplementation on microbial community composition have resulted in an increased relative abundance of *Bacteroideceae* and *Clostridia*. [17-19] To our knowledge, the CAZymes required for psyllium husk fermentation are not well defined yet. Surprisingly, no microbes were significantly enriched by inulin supplementation. This contradicts studies performed in humans, which found that inulin increases the relative abundance of *Bifidobacterium* and *Faecalibacterium*.[20]

Our final question asked if the variation between cohorts would impact the microbiome's response to fiber. We hypothesized the Moroccan cohort would have a greater microbiome diversity and fiber response when compared to the U.S. cohort because of conventionally higher dietary fiber intake. Research has demonstrated that non-industrialized populations usually have increased microbial and CAZyme diversity when compared to industrialized populations. [11] This may be due to reduced antibiotic exposure, differences in water treatment, or a higher intake of microbiota accessible carbohydrates, like dietary fiber. [11] Traditionally, higher gut microbiome diversity has been used as a measure of good gut health because it provides stability in the form of functional

redundancy as species compete to occupy similar niches. [12] Microbes competing within diverse environments are incentivized to streamline their genomes to occupy specialized niches and avoid overlapping competition, while microbes in environments with frequent perturbations, such as those caused by antibiotics, are incentivized to maintain larger genomes to maximize adaptability. [13] In agreement with this logic, we observed larger average genome sizes in the U.S. when compared to Moroccans. Unexpectedly, we observed significantly lower microbiome and CAZyme diversities in Moroccan versus U.S. samples at T0 (Figure 3.3A and 3D). One important caveat of our research is that there may be clinical or dietary influences we have not captured and could potentially explain if the Moroccan individuals we sampled live lifestyles that reduce gut microbiome diversity.

With respect to fiber response of each cohort, U.S. samples generally retained their taxonomic and CAZyme diversity at T24 more than when compared to Moroccan samples. U.S. samples displayed significantly decreased Shannon diversities when treated with inulin and pectin, while Moroccan samples exhibited a nonsignificant decrease in diversity in response to pectin. Decreased diversity suggests that the community became dominated by fewer microbes, which may be the desired outcome when using fiber to enrich for fiber degrading microbes. Despite the decreased overall diversity among Moroccan samples, there was a greater number of taxa enriched by pectin within these cultures. If the Moroccan cohort was more frequently exposed to pectin before sampling, this could explain their larger microbiome response. With respect to our U.S. samples, the typically higher presence of generalist taxa could explain why they responded to psyllium and pectin similarly, as measured by the number of enriched microbes.

When examining microbial enrichments of fecal communities *in vitro,* it is important to consider the culturing conditions. One important condition is the choice of growth medium. Here, we used brain-heart infusion (BHI) media, which is a rich media that does not select specifically for the growth of aero-intolerant microbes, unlike GAM or YCFA media. [21] Fermentation time is also important, as culturing selects for rapidly growing microorganisms and not necessarily primary fiber fermenters. We chose 24 hours due to its physiological relevance in the human gut, but our data suggests that many of the primary fermenters, such as *Bacteroides spp.*, were outcompeted by secondary fermenters, such as *Bifidobacteria,* as *Bacteroides spp.* was not significantly enriched. Thus, future experiments attempting to validate our results should be performed with additional time points, potentially in a chemostat, with a variety of growth mediums to allow the microbial communities to equilibrate.

In summary, this research demonstrates that fiber can selectively be used to promote the growth of specific organisms. Future directions involve characterizing the amount and composition of SCFAs to determine the health benefits of fiber supplementation. In the long term, further sequencing and analysis of fiber degrading microbial strains can be used to develop personalized fiber responses, paving the way for microbiome-mediated medicines.

METHODS

Fecal sample collection:

For U.S. samples, informed consent was obtained, and participants were given supplies and instructions for the self-directed collection of fecal samples. Samples were

returned anonymously and stored at -20°C. For the Moroccan cohort, the same procedure was followed, however, samples were transported in a refrigerated container until they reached the lab. Within 6 hours of the samples arriving in the lab, each sample was aliquoted into three 1.5ml tubes and maintained at -20°C.

Culturing experiments:

Culturing was performed by first thawing fecal samples on ice, then, 1g of fecal material was placed in a sterile, secondary tube and transferred to a Coy anaerobic chamber. Immediately, 5 mL of sterile, reduced PBS was added to each sample, and homogenized. Culturing was performed in 2 mL volumes in deep 96-well plates. Each well contained sterile, reduced BHI with either no fiber, apple pectin (Fisher, 15 g/L), inulin from chicory (Fisher, 15 g/L), or psyllium husk powder (Now Foods, 8 g/L). Pectin media was pH neutralized using NaOH. Wells were inoculated with 40 uL of fecal slurry. Controls consisted of BHI with no fiber and no inoculum, and BHI with each fiber but no inoculum. Afterwards, cultures were mixed by pipetting and a 1 mL aliquot for OD600 was taken and stored at -80°C. Plates were sealed with a silicone lid and allowed to incubate for 24 hours, after which they were stored at -80°C. Microbial growth was verified by taking OD600 measurements before and after 24 hours.

DNA extraction:

For our initial time point, 200 mg of fecal material was thawed on ice and was extracted using ZymoBiomics DNA Miniprep Kit (Cat. #D4300) according to the manufacturer's protocol. For our 24-hour time point, cultures were thawed on ice and 250 uL were subjected to DNA extraction according to the manufacturer's protocol. Bead lysis

during the extraction was performed at 6.5 m/s for 5 minutes total (MPBio FastPrep-24). A mock community standard was included as a positive extraction control (Cat. #D6300). Water was used for negative extraction controls.

Shotgun library preparation and sequencing:

Libraries for shotgun sequencing were prepared using the Illumina DNA prep kit (Cat. # 20018705), using our published low-volume protocol which reduces input DNA and enzyme volumes tenfold. [22] Five microliters or 50 ng (whichever was reached first) of DNA per sample was tagmented according to our adapted protocol. Afterwards, i5 and i7 indices were added to each sample in 1.25 uL volumes and annealed via PCR using 10 uL of KAPA HiFi HotStart ReadyMix (Cat. # 7958935001). Libraries were pooled then size-selected and cleaned using 56 and 14.4 uL of the included sample purification beads, respectively. Positive and negative sequencing controls included the ZymoBIOMICS Microbial Community DNA Standard (Cat. #D6305) and purified water. The pooled library's quantity was assessed using a Qubit Fluorometer and the Quanti-iT PicoGreen dsDNA kit (Cat. #P7589). Fragment size was checked on an Agilent Tapestation using an Agilent Bioanalyzer High Sensitivity DNA Analysis kit (Cat. #5067-4626). Lastly, the library pool was sequenced on an Illumina Novaseq 6000 using an S4 flow cell at the Genomics High Throughput Facility at the University of California Irvine. This produced an average of 3,061,344 +/- 2,213,848 (σ) paired-end reads per sample, 150 base-pairs in length.

MAG table generation:

Raw sequences first had sequencing adapters, artifacts, and low-quality sequences removed using the BBMap v38.79 script 'bbduk.sh' with the parameters "ref=adapters,

artifacts, phix, lambda, pjet, mtst, kapaq, trim=w, trimq=24, forcetrimleft=15". Next, duplicate reads were removed with the 'dedupe.sh' script from BBMap using the default settings. Human-derived reads were removed using BowTie2 v2.4.5 using the default parameters and hg38 as the reference genome. Subsequently, quality-filtered samples were cross assembled using MEGAHIT and the parameters "--presets meta-large --min-contig-len 2500". From the cross-assembled contigs, MAGs were binned using MetaBAT with default parameters. Quality of MAGs was assessed with CheckM using the default parameters and dereplication of MAGs was done using dRep with the default settings. Taxonomic assignment of dereplicated MAGs was done using GTDB-Tk with the "--classify_wf" preset. MAGs were concatenated, and BowTie2 was used to build an index and align per sample sequences to the index. Lastly, the 'pileup.sh' script from BBMap was used to combine each samples alignment to specific MAGs. This produced a table of annotated MAGs and the number of read counts per sample.

Annotation of carbohydrate active enzymes:

CAZyme annotation was performed twice. The first method was used to assess the community wide CAZyme composition, while the second method was used to investigate the CAZymes encoded within each MAG. The community-wide CAZyme annotation was performed on cross-assembled contigs. For this, open reading frames (ORFs) were assigned with Prodigal v2.6.3 and then annotated with dbCAN2 using the default parameters for both programs. A table of per sample annotated ORF counts was obtained with the same methods as for the MAG table generation. First, BowTie2 was used to make an index of annotated contigs, then individual samples were aligned to the index. 'Pileup.sh'

was used to compile the results into a singular table. Lastly, per sample ORF counts were normalized to reads per kilobase per genome equivalent using MicrobeCensus v1.1.1 on default parameters. The second CAZyme annotation occurred at the MAG level. Here, we ran each non-dereplicated MAG through prodigal and dbCAN2. The resulting annotations were imported in R for analysis afterwards.

Data analysis:

Analysis of MAGs and CAZymes was performed in R v4.2.1. The Vegan v2.6-2 package was used to calculate Shannon diversity with the 'diversity' function, PERMANOVA with the 'adonis2' function, NMDS with the 'metaMDS' function. Bray-Curtis dissimilarity was used as the distance metric where possible. Significance testing of normally distributed data, such as Shannon diversity, was performed using linear-mixed effect models with the nlme v3.1-159 package. Differential abundance of MAGs was determined using the Maaslin2 package in R. Pangenomes were visualized and annotated using Anvi'o v7.1.

Volatile extraction and gas chromatography mass spectrometry

Volatile extraction began by taking 200µL from each culture replicate (three replicates total) and combining them into a 1.5 mL microcentrifuge tube for a total of 600 µL. 500 µL of the 600µL pooled sample was transferred to a separate microcentrifuge tube, and 0.16 µL of 7.8 M 2-ethylbutyric acid was added to create a 2.5 mM 2-ethylbutyric acid internal standard. 150 µL of the fecal sample was transferred into three separate vials, where a cap, lid liner and sorbent pen for headspace analysis was quickly placed onto the vials to prevent the loss of volatile compounds. Vials were placed under vacuum and transferred over to a 5600 SPEU shaking incubator. Volatiles were extracted at 70℃ for 1

hour at 200 rpm. After extraction, vials were placed on a cold block for 15 minutes. Samples were run on an Agilent 7890A GC 5975C MS at 260 ℃ for 38 minutes.

ACKNOWLEDGEMENTS: We would like to express our gratitude to Soumaya El Halas for collecting the Moroccan samples, making this work possible. We would also like to thank the Fulbright Scholar Program, Graduate Assistance in the Areas of National Need Fellowship, Rose Hills Foundation Fellowship, and the Northarvest Dry Beans Foundation for funding this research.

REFERENCES

1. Wang, M. *et al.* In vitro colonic fermentation of dietary fibers: Fermentation rate, short-chain fatty acid production and changes in microbiota. *Trends in Food Science & Technology* 88, 1–9 (2019).

2. Makki, K., Deehan, E. C., Walter, J. & Bäckhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host & Microbe* 23, 705–715 (2018).

3. Sonnenburg, J. L. & Sonnenburg, E. D. Vulnerability of the industrialized microbiota. *Science* 366, eaaw9255 (2019).

4. Myhrstad, M. C. W., Tunsjø, H., Charnock, C. & Telle-Hansen, V. H. Dietary Fiber, Gut Microbiota, and Metabolic Regulation—Current Status in Human Randomized Trials. *Nutrients* 12, 859 (2020).

5. Cantu-Jungles, T. M. & Hamaker, B. R. New View on Dietary Fiber Selection for Predictable Shifts in Gut Microbiota. *mBio* 11, e02179-19 (2020).

6. Elshahed, M. S., Miron, A., Aprotosoaie, A. C. & Farag, M. A. Pectin in diet: Interactions with the human microbiome, role in gut homeostasis, and nutrient-drug interactions. *Carbohydrate Polymers* 255, 117388 (2021).

7. Ndeh, D. *et al.* Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature* 544, 65–70 (2017).

8. Scott, K. P., Martin, J. C., Duncan, S. H. & Flint, H. J. Prebiotic stimulation of human colonic butyrate-producing bacteria and bifidobacteria, *in vitro*. *FEMS Microbiol Ecology* 87, 30–40 (2014).

9. Roberfroid, M. B. Prebiotics and probiotics: are they functional foods? *The American Journal of Clinical Nutrition* 71, 1682S-1687S (2000).

10. Hussain, M. A., Muhammad, G., Jantan, I. & Bukhari, S. N. A. Psyllium Arabinoxylan: A Versatile Biomaterial for Potential Medicinal and Pharmaceutical Applications. *Polymer Reviews* 56, 1–30 (2016).

11. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and implications for human health. *Nat Rev Microbiol* 17, 383–390 (2019).

12. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* 350, 663–666 (2015).

13. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. *Nat Commun* 8, 1162 (2017).

14. Lopez-Siles, M. *et al.* Cultured Representatives of Two Major Phylogroups of Human Colonic Faecalibacterium prausnitzii Can Utilize Pectin, Uronic Acids, and Host-Derived Substrates for Growth. *Appl Environ Microbiol* 78, 420–428 (2012).

15. Yang, J., Martínez, I., Walter, J., Keshavarzian, A. & Rose, D. J. In vitro characterization of the impact of selected dietary fibers on fecal microbiota composition and short chain fatty acid production. *Anaerobe* 23, 74–81 (2013).

16. Reichardt, N. *et al.* Specific substrate-driven changes in human faecal microbiota composition contrast with functional redundancy in short-chain fatty acid production. *ISME J* 12, 610–622 (2018).

17. Shulman, R. J. *et al.* Psyllium Fiber Reduces Abdominal Pain in Children With Irritable Bowel Syndrome in a Randomized, Double-Blind Trial. *Clinical Gastroenterology and Hepatology* 15, 712-719.e4 (2017).

18. Yang, C. *et al.* The effects of psyllium husk on gut microbiota composition and function in chronically constipated women of reproductive age using 16S rRNA gene sequencing analysis. *Aging* 13, 15366–15383 (2021).

19. Gamage, H. K. A. H. *et al.* Fiber Supplements Derived From Sugarcane Stem, Wheat Dextrin and Psyllium Husk Have Different In Vitro Effects on the Human Gut Microbiota. *Front. Microbiol.* 9, 1618 (2018).

20. Le Bastard, Q. *et al.* The effects of inulin on gut microbial composition: a systematic review of evidence from human studies. *Eur J Clin Microbiol Infect Dis* 39, 403–413 (2020).

21. Tidjani Alou, M. *et al.* State of the Art in the Culture of the Human Microbiota: New Interests and Strategies. *Clin Microbiol Rev* 34, e00129-19 (2020).

22. Weihe, C. & Avelar-Barragan, J. Next generation shotgun library preparation for Illumina sequencing - low volume v1. doi:10.17504/protocols.io.bvv8n69w

23. Gündüz Ergün, B. & Çalık, P. Lignocellulose degrading extremozymes produced by Pichia pastoris: current status and future prospects. Bioprocess Biosyst Eng 39, 1–36 (2016).

24. Martins, L. C., Monteiro, C. C., Semedo, P. M. & Sá-Correia, I. Valorisation of pectin-rich agro-industrial residues by yeasts: potential and challenges. Appl Microbiol Biotechnol 104, 6527–6547 (2020).

TABLES AND FIGURES

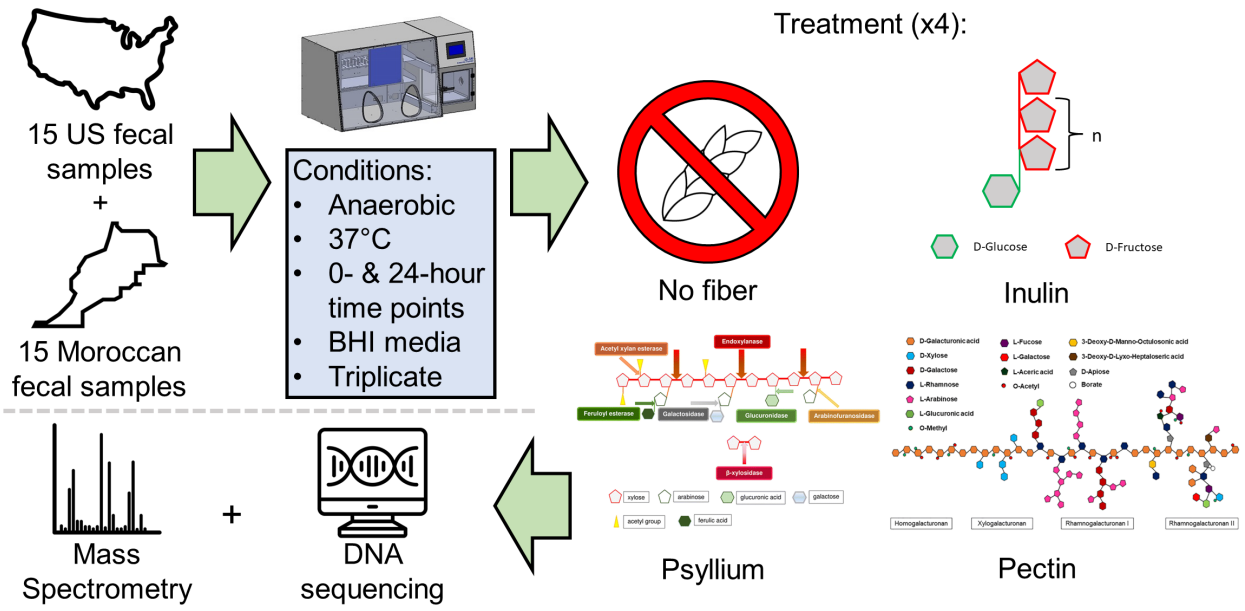


Figure 3.1: Study design. Thirty fecal samples from different U.S. and Moroccan individuals were collected for this study. The feces were used to inoculate anaerobic liquid cultures filled with BHI and allowed to ferment for 24 hours at 37C. Cultures were supplemented with either no fiber, inulin (15g/L), pectin (15 g/L), or psyllium husk powder (8 g/L). Microbial growth was verified using optical densities before and after fermentation. Afterwards, cultures had their DNA extracted, and sequenced using shotgun metagenomic sequencing. Gas-chromatography mass spectrometry was also performed to measure the abundance of short chain fatty acids and other volatile compounds. The structural diagrams for psyllium and pectin are from Gündüz Ergün *et al.* and Martins *et al.,* respectively. [23, 24]

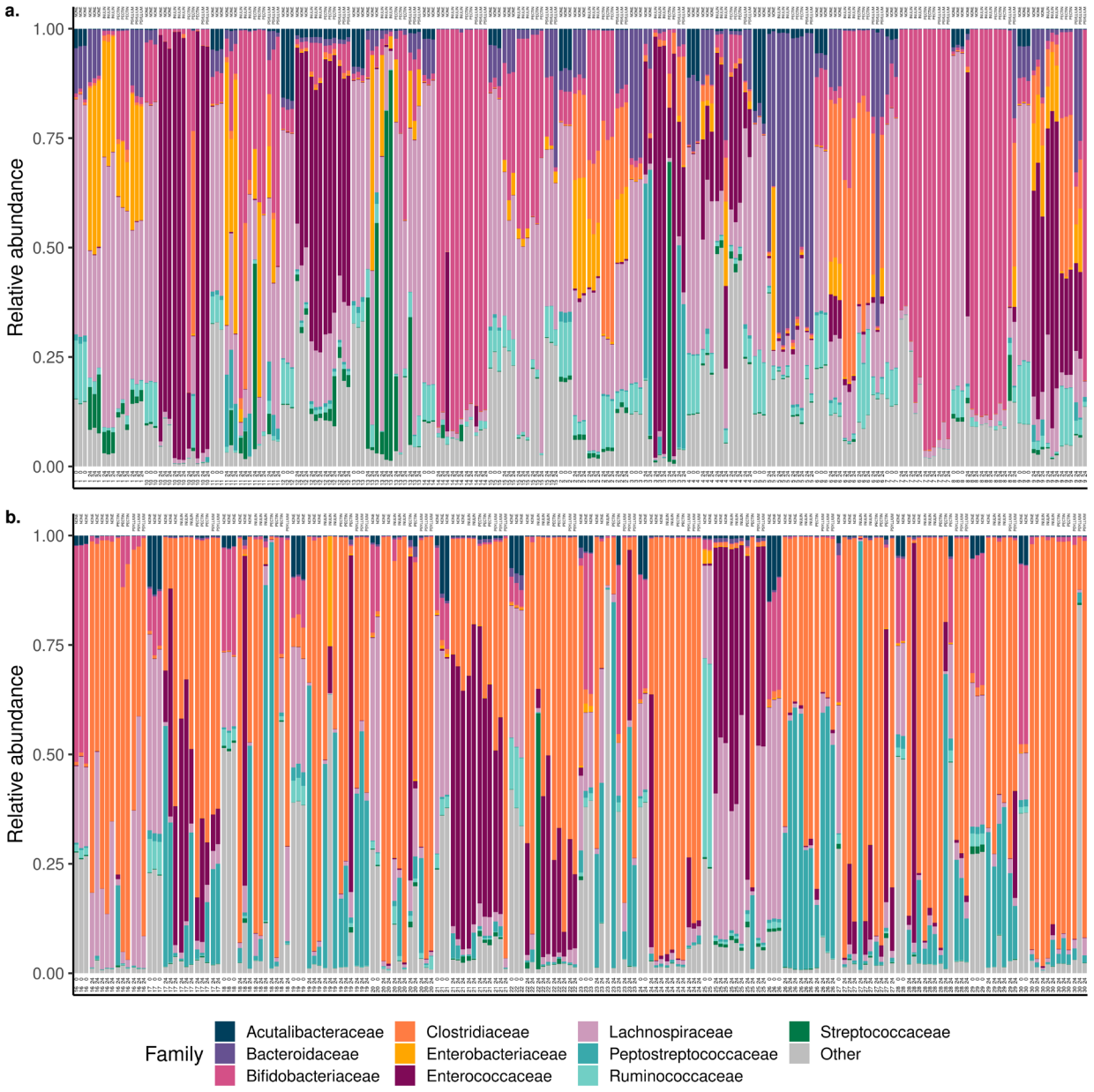


Figure 3.2: Taxonomic characterization of fecal community cultures. A stacked bar plot displaying the relative abundance of microbes at the family level for all samples. Each bar is one sample, with multiple samples per time point, treatment, and individual. Panels **a.)** and **b.)** correspond to samples from the U.S. and Morocco, respectively. The text in the bottom margin refers to the time point (top) and subject ID (bottom).

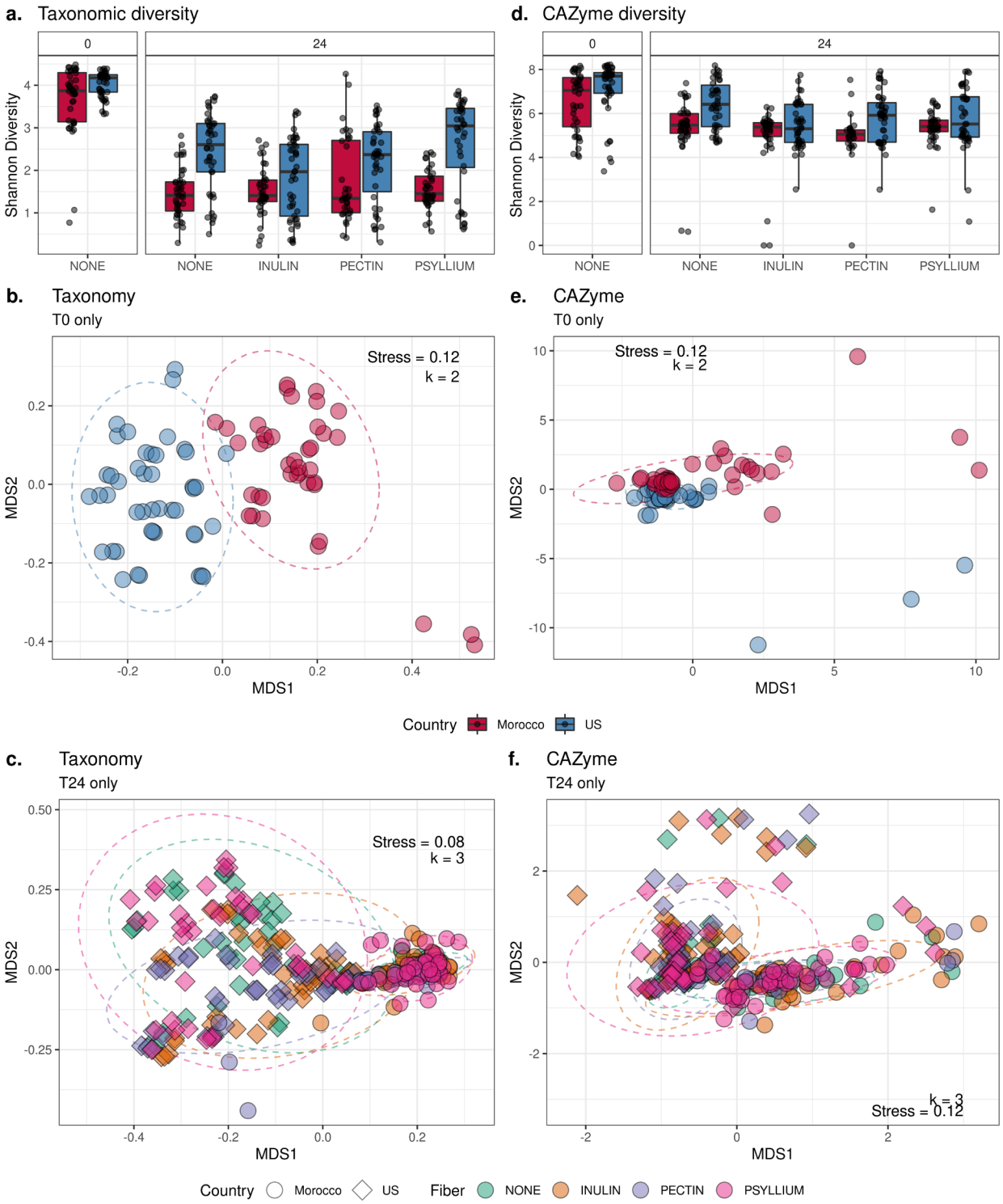


Figure 3.3: Country, time, and fiber treatment are all significantly associated with microbial and enzymatic diversity and composition. **a.** & **d.)** Box and whisker plots

illustrating the Shannon diversity of taxa **(a.)** & carbohydrate active enzymes **(d.).** Samples are colored by country and faceted by time. **b. & e.)** Nonmetric multidimensional scaling ordination performed on the Bray-Curtis dissimilarity matrices of samples using their taxonomic **(b.)** or CAZyme abundances **(e.)** at time point 0 without the addition of fiber. Points are colored by country, and a 95% confidence interval surrounds each cohort. **c. & f.)** Nonmetric multidimensional scaling ordination performed on the Bray-Cutris dissimilarity matrices of samples using their taxonomic **(c.)** or CAZyme abundances **(f.)** at time point 24h with the addition of fiber. Points are colored by fiber, shaped by country, and a 95% confidence interval surrounds each cohort by fiber. Across all figures, each point represents a single replicate.

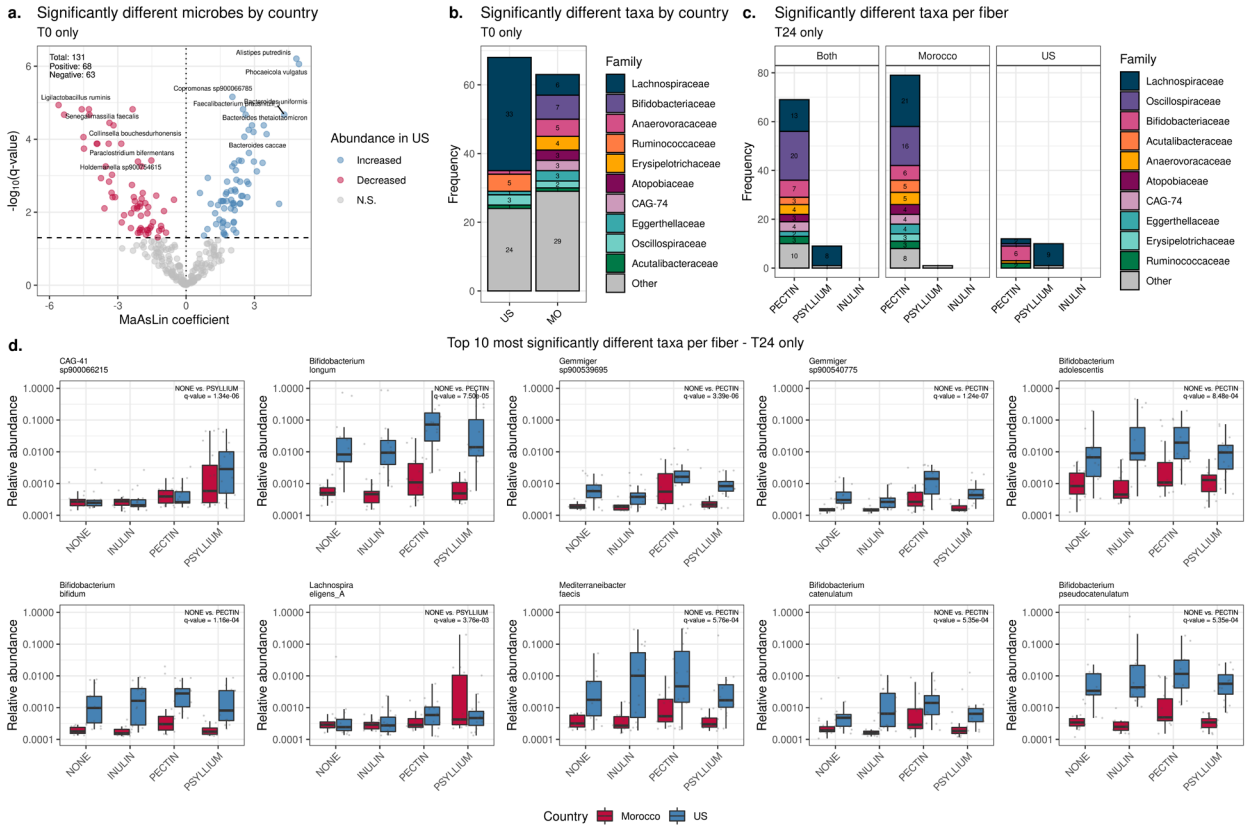


Figure 3.4: Pectin and psyllium enrich for microbes, but inulin does not. **A.)** A volcano plot displaying significantly different microbes between countries at time point 0h. The vertical and horizontal lines denote a zero-fold change and a q-value of less than 0.05, respectively. Points are colored by their abundance, with blue meaning that the microbe was significantly more abundant in the U.S. and red meaning that the microbe was significantly less abundant in the U.S., or more abundant in Moroccans. **B.)** A stacked bar plot summarizing the frequency and taxonomy, at the family level, of microbes determined to be significantly different in abundance between countries at 0h. **C.)** A stacked bar plot summarizing the frequency and taxonomy, at the family level, of microbes determined to be significantly enriched by fiber when compared to untreated controls at 24h. Differential abundance testing was performed with all samples together, or within each cohort. **D.)** Box

and whisker plots showing the relative abundance of the top 10 taxa determined to be the most significantly enriched by fiber when compared to controls.

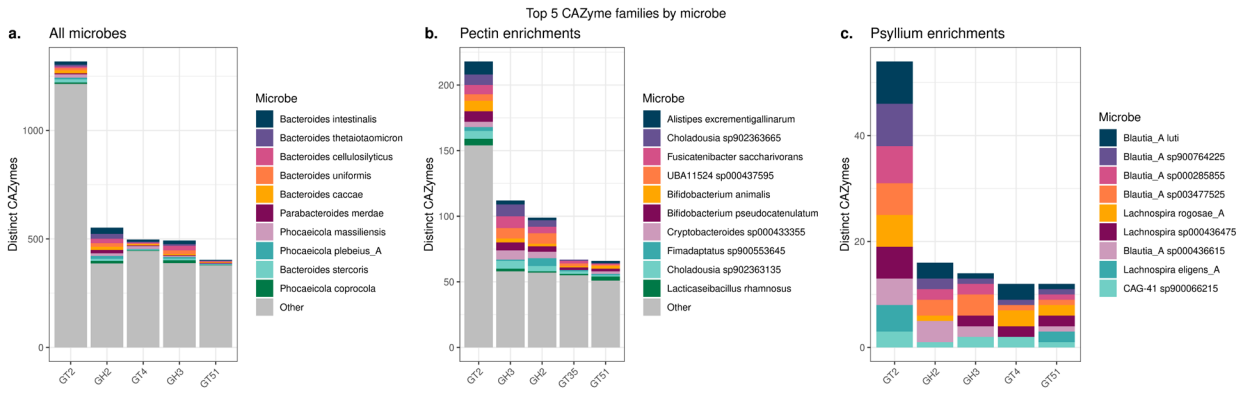


Figure 3.5: Characterization of microbial Carbohydrate Active Enzymes. Stacked bar plots showing the number of distinct CAZymes per CAZyme family. Only the top 5 most abundant CAZyme families are shown. Panel **a.)** corresponds to all microbes, **b.)** corresponds to microbes only significantly enriched by pectin when compared to no fiber controls, and **c.)** corresponds to microbes only significantly enriched by psyllium when compared to no fiber controls.

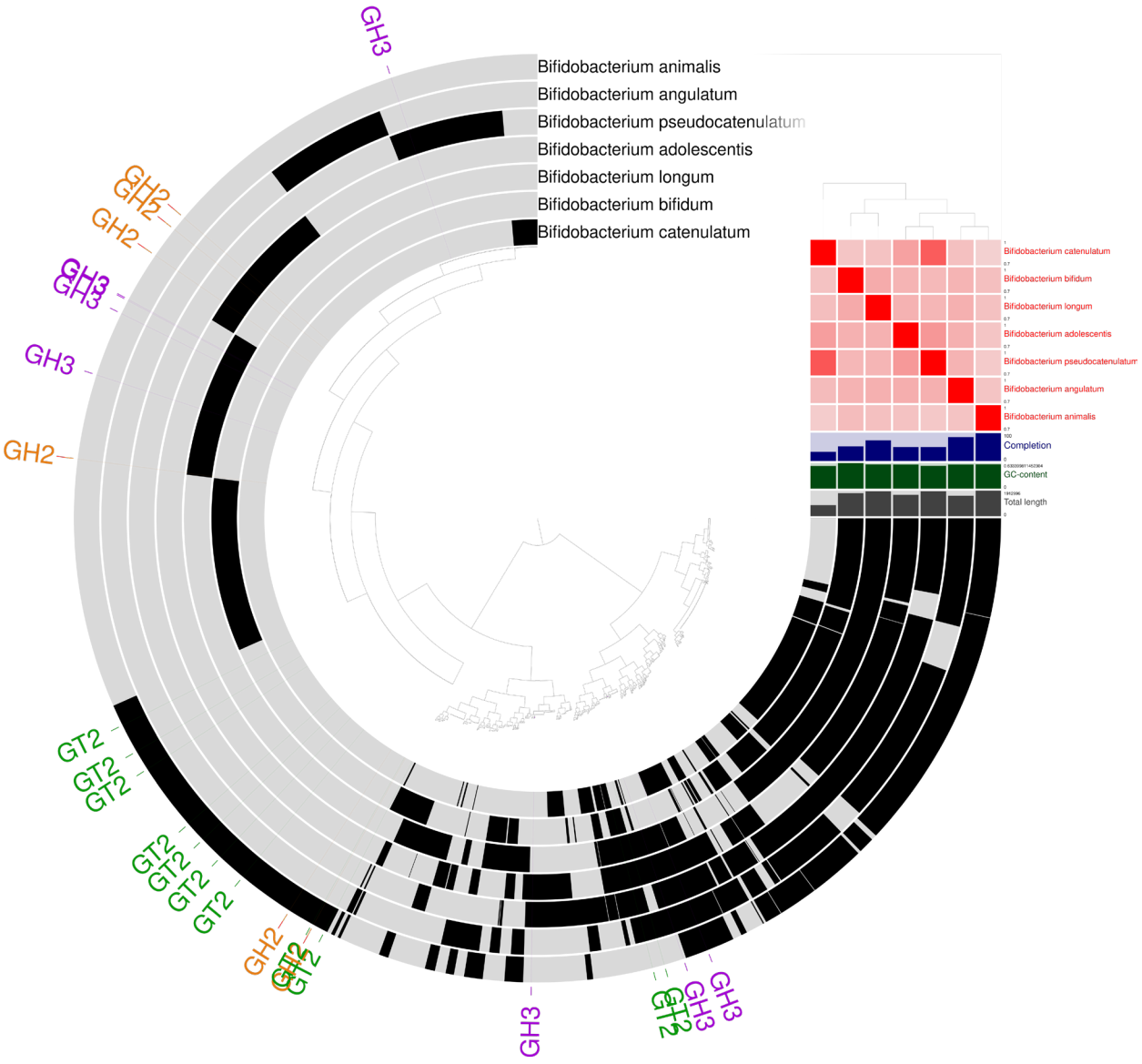


Figure 3.6: Characterization of microbial Carbohydrate Active Enzymes within Bifidobacterium pangenomes. A circular dendrogram displaying the pangenome of 7 *Bifidobacterium* species. Above each row is the total genome length in bases, the GC content, genome completion, and average nucleotide identity.

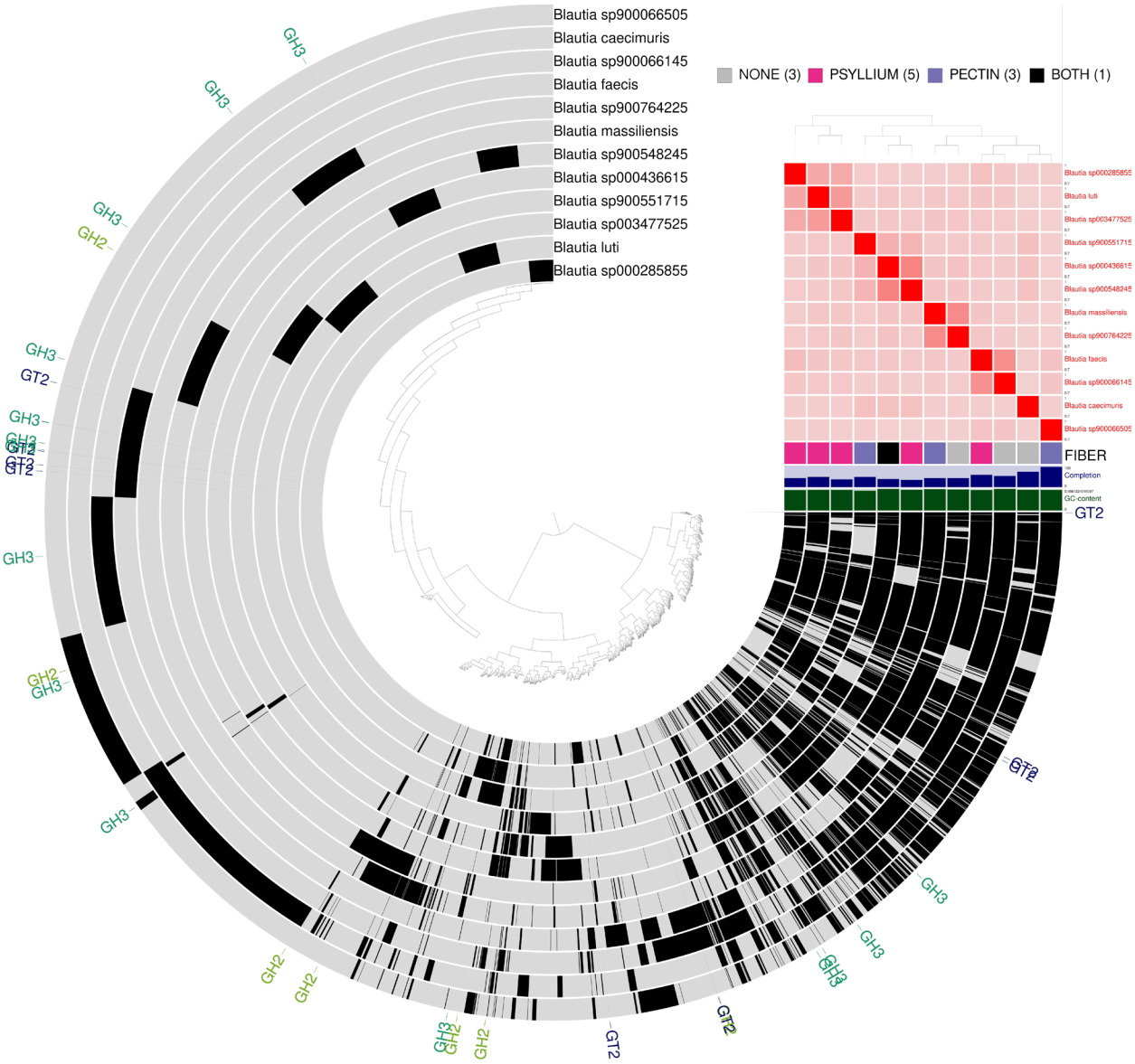
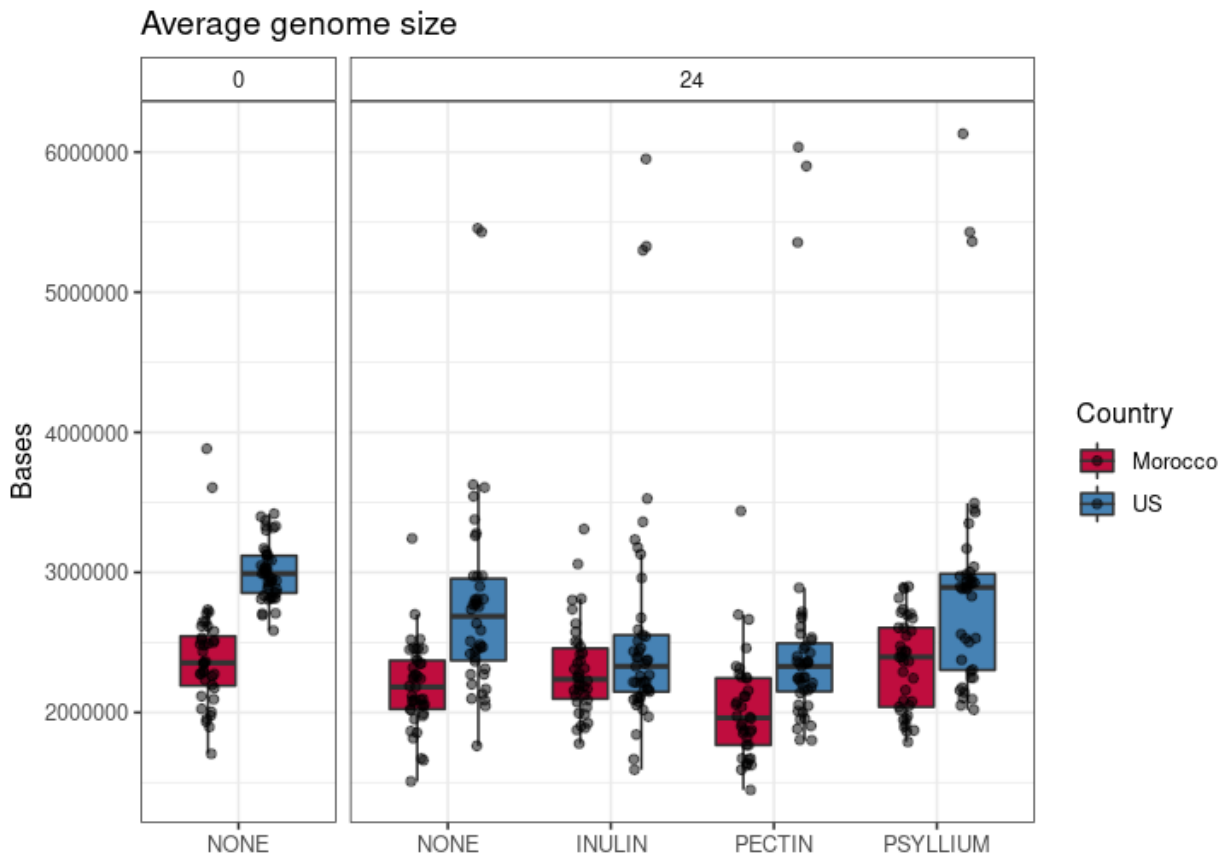
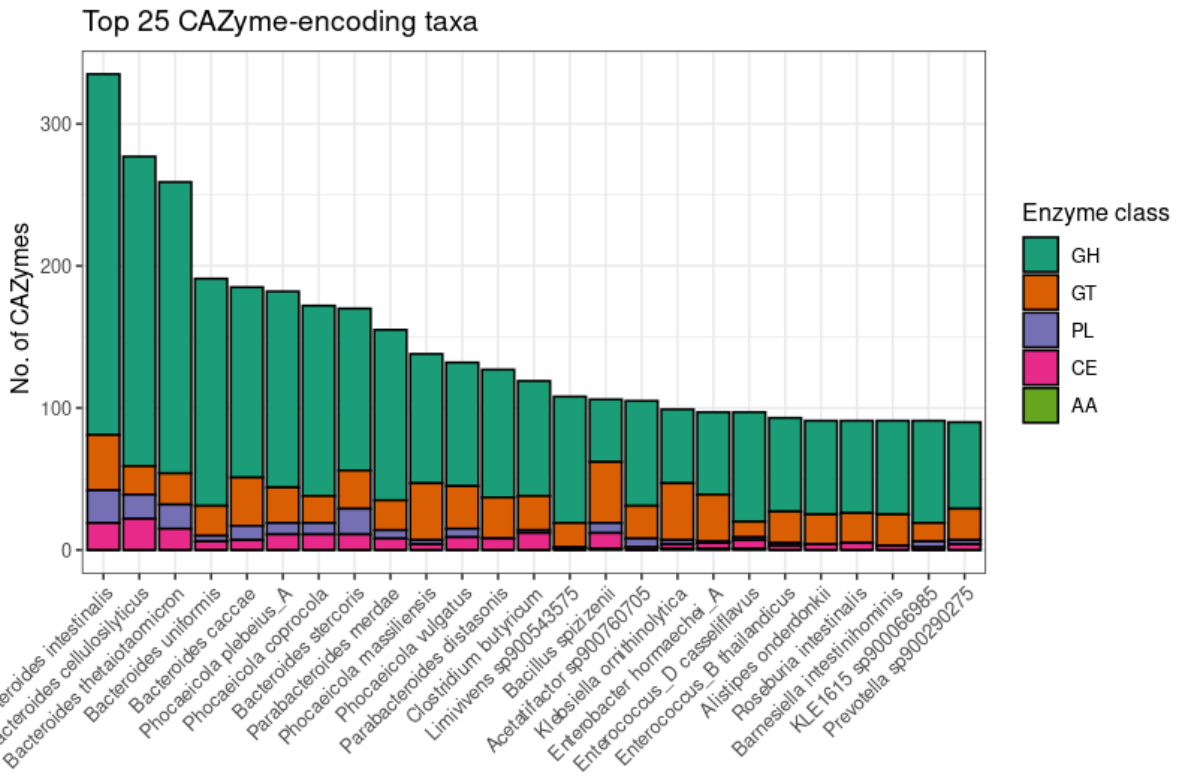


Figure 3.7: Characterization of microbial Carbohydrate Active Enzymes within Blautia pangenomes. A circular dendrogram displaying the pangenome of 12 *Blautia* species. Above each row is the GC content, genome completion, whether the species was significantly enriched by fiber, and average nucleotide identity.



Supplementary Figure 3.1: Average genome sizes per sample. Box and whisker plots displaying the average genome size, in bases, of all microbes in each sample. Samples are faceted by time, fiber treatment, and country.



Supplementary* Figure 3.2: *Number of CAZymes per microbe. A stacked bar plot displayed the number of unique CAZymes encoded by each microbe. Each bar is colored by the type of enzyme class found within the carbohydrate active enzymes. GH = glycoside hydrolase, GT = glycosyl transferase, PL = polysaccharide lyase, CE = carbohydrate esterase, AA = Auxiliary activities.

SUMMARY AND FUTURE DIRECTIONS

In my thesis, I have shown the following novel discoveries: 1) The composition of the mucosal gut microbiome is distinct between the adenoma-carcinoma sequence and the serrated pathway of colorectal carcinogenesis. This information could be used to accurately predict the polyp type of a sample. One of the hallmark features enabling this prediction was the depletion of *Eggerthella lenta* in samples originating from serrated polyp samples. Lastly, I characterized a novel method of microbiome sampling and compared it to conventional fecal sampling. 2) The gut microbiome of patients with myeloproliferative neoplasms differs based on the subtype of the disease. Specifically, individuals with myelofibrosis had significantly reduced microbial diversity and higher beta-dispersion when compared to individuals with polycythemia vera and essential thrombocythemia. The altered microbiome of subjects with myelofibrosis significantly correlated with elevated inflammation. I also demonstrate that a 10-week active dietary intervention based on counseling to follow a Mediterranean diet does not significantly alter the gut microbiome. 3) Finally, I demonstrate with *in vitro* fecal community culturing that pectin and psyllium can be used to enrich for microbes across two different cohorts. The quantity and taxonomy of microbes was dependent on the type of fiber used, which correlated with the structural complexity of each fiber.

In Chapter 1, I described signatures of the microbiome which suggested that the lack of dietary fiber could potentially explain the development of colorectal cancer via the serrated pathway. This pathway is characterized by aberrant epigenetic expression of genes that promote carcinogenesis, and butyrate has roles in host epigenetic regulation.

Future studies could continue to characterize the serrated microbiome using mucosal sampling techniques, while incorporating dietary metadata and fecal butyrate quantification. Studies aiming to elucidate the role of the serrated microbiome on host epigenetic regulation could perform *in vitro* or *in vivo* experiments which involve bisulfite sequencing to quantify any differences in host genome methylation produced by the microbiome. For example, one could take the gut microbiome of a healthy individual, a serrated-polyp bearing individual, and a tubular adenoma-bearing individual and transplant it into a germ-free colorectal cancer mouse model and measure the amount of methylation. In this scenario, I hypothesize that the serrated polyp associated microbiome would promote more aberrant epigenetic events.

In chapter 2, I demonstrated that there were no changes in the gut microbiome associated with the Mediterranean diet intervention. Though we characterized the gut microbiome of these individuals, the main purpose of this study was to see if individuals with MPN could adhere to a Mediterranean diet. In the future, I would like to investigate the gut microbiome of MPN patients who receive a 6-month Mediterranean diet intervention. In an ideal environment, meals would be provided to strengthen the effect of the intervention. I hypothesize that this would produce a reduction in inflammation in individuals with MPN, and that this reduction would be associated with changes in the gut microbiome.

Finally, in chapter 3, I have shown that fiber can enrich for specific microbes. One interesting result has been the dominance of *Clostridia* in Moroccan samples, regardless of fiber treatment. U.S. samples also have *Clostridia,* but often become dominated by other

microorganisms. Moving forward, I would like to characterize the strain level differences between the *Clostridia* and *Bacteroides* of both cohorts. Additionally, I want to continue to explore the CAZyme profiles to compare the levels of CAZymes implicated in pectin and psyllium degradation across cohorts, which would help identify novel pectolytic microbes. Finally, I would like to quantify and correlate short-chain fatty acid abundances with microbial relative abundances across fiber treatments and cohorts at 24 hours. Here, I hypothesize that inulin would produce the greatest amount of short-chain fatty acids of the three fibers because of its structural simplicity and potentially greater accessibility to microbial fermentation. Together, this research would assist in enabling the use of prebiotics to promote a personalized microbiome response in an effort to combat disease associated with gut microbiome dysbiosis and chronic inflammation.