NEXT-GENERATION SEQUENCING BASED QUANTIFICATION OF MICROBIAL COMMUNITIES AND GENE PATHWAYS

by

Roshonda Barner Jones

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2016

Approved by:

____
Dr. Anthony Fodor

____
Dr. Wei Sha

____
Dr. Zhengchang Su

____
Dr. Xinghua Shi

____
Dr. Christine Richardson

ABSTRACT

ROSHONDA BARNER JONES. Next-generation sequencing based quantification of microbial communities and gene pathways. (Under the direction of DR. ANTHONY A. FODOR)

Humans act as a host to trillions of microorganisms. The collective of these microorganisms is called a host’s microbiota. There is a growing interest in the bacterial composition of the gut microbiome of humans (structure) and the role of the microbiome in human diseases (function). Many methods are used to define the structure and the function of a microbial community and there are concerns about how the use of these varying methods can impact the reproducibility of microbiome research. In this dissertation, we aimed to determine how different measurement techniques impacts the understanding of the structure and function of the gut bacterial communities in humans and in model systems such as non-human primates and rodents. Along with determining these different techniques to quantify the structure and function of the gut microbiome, this dissertation applies these different techniques to determine how dietary sugars impact microbial community composition and determine factors that are associated with the gut microbiome in patients with colorectal adenomas, a benign tumor which is often times a precursor to colorectal cancer.

DEDICATION

This dissertation is dedicated to my family especially my lovely mother Rosalind Barner and my dear husband Brannon Jones who have supported me throughout this process.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Bill and Melinda Gates Foundation for their generous financial support. I would like to thank my dissertation committee chair, Dr. Anthony Fodor, my dissertation committee members Dr. Wei Sha, Dr. Xinghua Mindy Shi, Dr. Zhengchang Su, and Dr. Christine Richardson. I would also like to thank those who collaborations made this research possible including Dr. Kylie Kavanagh of Wake Forest University, Dr. Martha Shrubsole of Vanderbilt University, Dr. Michael Goran of University of Southern California and Dr. Temitope Keku of University of North Carolina – Chapel Hill.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

16S	16S small subunit ribosomal RNA gene
AbundantOTU	Tool used to create consensus sequences from 16S sequence reads
ANOVA	Analysis of Variance
BH	Benjamini-Hochberg Correction
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
CRC	Colorectal Cancer
DNA	Deoxyribonucleic Acid
FDR	False Discovery Rate
HFCS	High-Fructose Corn Syrup
HMP	Human Microbiome Project
HS	Hepatic Steatosis
HUMAnN	HMP Unified Metabolic Analysis Network
IBD	Inflammatory Bowel Disease
IL	Interleukin
KEGG	Kyoto Encyclopedia of Genes and Genomes
MALDI-TOF	Matrix-Assisted Laser Desorption Ionization Time-Of-Flight
MDS	Multidimensional Scaling
MS	Mass Spectrometry
MT	Microbial Translocation
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction

PICRUSt	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
PPCCT	Personalized Prevention of Colorectal Cancer Trial
QIIME	Quantitative Insights Into Microbial Ecology
R	GNU Project Language and Environment for Statistical Computing
RDP	Ribosomal Database Classifier
RNA	Ribonucleic Acid
RNA-Seq	RNA sequencing
SSB	Sugar Sweetened Beverage
UC	Ulcerative Colitis
UCLUST	Algorithm used to cluster sequences based on sequence similarity
WGS	Whole-Genome Shotgun
ZIG	Zero-inflated Gaussian

CHAPTER 1: MEASUREMENT TECHNIQUES USED IN MICROBIOME RESEARCH

1.1 Introduction

Animals act as a host to trillions of microorganisms. The collective of these microorganisms is called a host's microbiota. While the human genome has about 26,000 genes, the bacterial community of the human gut is predicted to have close to 3 million genes[1]. From this fact, one can assume that all of the functions that the body needs to maintain itself are not strictly coded in the human genome but that the collective genomes of the microbiota that we house, namely our microbiome, is involved in carrying out some of these functions. Some examples of this symbiotic relationship include but are not limited to: immune protection[2]; digestion[3]; and protection from pathogens[4, 5].

There are many factors that can shape the microbial community of an animal host including the site of the microbial community on the body[6], age of host[7-10], the host's diet[11-13], and disease state of the host[2, 14-17]. Disease can play a role because it is associated with dysbiosis or a disruption of the microbial-host relationship. Dysbiosis has been correlated with a number of conditions such as inflammatory bowel disease (IBD)[11], obesity[18], diabetes[19], and many others. The host's housing environment has also been shown to have an impact on their microbial community[20, 21]. Examples of this phenomenon include elderly humans whose microbiota differs based on where they live[9] and mice whose microbiota differs based on the cage that the mouse is housed[22].

In microbiome research, the questions that we often ask are: What bacteria are present in a community and how the composition of these communities change and what these microbes are functionally capable of doing. To determine which bacteria are present in a community we often look to the 16S small subunit ribosomal RNA gene (16S). This marker gene is useful because it is composed of regions that are variable from species to species and 16S sequence reads can be classified into clusters based on sequence similarity. Other methods that are used to elucidate these bacterial communities include whole-genome metagenome shotgun (WGS) sequencing in which all DNA in a sample is sequenced and RNA-seq which measure the transcriptome of a system at a given time.

Research in the field of metagenomics and microbiome studies can be complex There are a large number of methods used to define the structure and the function of a microbial community and there are concerns about how these methods impact the reproducibility of microbiome research[23]. Some of the factors that impact the outcome of an analysis are the type of sequences that are used such as 16S gene sequences versus sequences generated using WGS sequencing. These sequences types can differ such that 16S rRNA marker gene sequence reads primarily gives insight about what *bacteria* are in a community and WGS sequence gives insight about what *genes* are in a community. However, there have been cases in which 16S rRNA sequences are used to predict the gene profile of a microbial community and WGS sequences reads are normally used to give insight about the composition of a community [24, 25].

1.2 Problem Statement

In this work we aim to determine how different measurement techniques impact our understanding of the structure and function of the gut bacterial communities in humans, primate and rodent model systems. Along with determining these different techniques to quantify the structure and function of the gut microbiome, this work will apply these different techniques to determine how dietary sugars impact microbial community composition and determine factors that are associated with the gut microbiome in patients with colorectal adenomas.

1.3 Significance

One goal of this dissertation is to compare a number of different ways that microbial communities can be studied. We first compare whole-genome sequencing, 16S rRNA sequencing techniques and methods that simulate whole-genome from 16S sequence reads. We then compare the quantification of gut microbial communities using stool samples versus swab samples. This is of great interest because collecting stool samples is much more prevalent in gut microbiome research in part because stool samples are an easily collectable source of ample microbial DNA [26]. However, swab samples have also been used[27] and may in some cases be easier to collect than stool samples so it would be beneficial to determine the differences in microbial quantifications from these sample types.

Our second overall goal involves a biological applications of the technical considerations of our first goal. Our first biological question involves the role of dietary sugar in shaping the microbial community. Increased sugar consumption is linked to the obesity epidemic that has swept the United States and much of the world over the last three decades. Obesity can lead to a number of other conditions such as diabetes, high blood pressure and high cholesterol [28]. It is believed that the increase use of fructose, mainly high-fructose corn syrup, in our diets is a major contributor to this epidemic[29]. Increased amounts of high-fructose corn syrup have been introduced into our diets over the last several decades through increased consumption of soft drinks[30]. It has been shown that rats that are fed high-fructose corn syrup exhibit signs of obesity[31]. We aim to determine if sugar in general, and fructose in particular, has an impact on the gut microbial makeup of these rats, and towards this aim we have 16S sequences collected from rats fed a sugar solution and those on a control diet.

Lastly, we aim to determine how changes in the microbial community are associated with the presence of colorectal adenomas in multiple studies. This is important because patients with adenomas have a high risk of developing colorectal cancer. Finding any microbial associations with colorectal adenomas could allow for creation of a new non-invasive diagnostic to determine whether or not a patient has colorectal adenomas and should be further screened for the presence of colorectal cancer. It is also as important to determine if these microbial associations are reproducible across multiple cohorts.

1.4 Research Objectives and Approaches

The four research questions that we address in this work are summarized as follows:

Aim 1. Can swab samples adequately replace stool samples when analyzing the microbial community of the human gut?

Aim 2. How well does the functional profile of a microbial community predicted by 16S marker gene sequences match the functional profile of a microbial community calculated with whole-genome shotgun metagenome sequences on a non-human primate system? Does the amount of fructose in diets of primates impact the gut microbial communities of non-human primates?

Aim 3. How does fructose amount impact the microbial community of rats?

Aim 4. Are the microbes that are associated with colorectal adenomas in one study reproducible in other studies?

1.5 Statistical Approach

Throughout this dissertation we will employ linear models as an approach to describe aspects of our data. There are a number of assumptions that these models make about the data including the assumptions of normality, homogeneity, and independence, among others. The violation of the assumption of independence is said to be the most dangerous violation because tests that violate this assumption tends to produce many false positives[32]. A violation of independence occurs when values of certain explanatory variables influence each other. An example of this is when longitudinal samples are taken from multiple subjects but then these samples are treated as if they were sampled from different individuals. Since violating this assumption can be a major source of error in inference we will often utilize marginal linear models, which relaxes the assumptions of independence.

Marginal linear models follow the standard linear model $Y_i = X_i \times \beta + \varepsilon_i$ such that Y_i is normally distributed with a mean of $X_i \times \beta$ and a variance of V_i for each group i. Examples of groups include subjects, hospitals, or the cages that subjects are housed for any study. The simplest scenario for V_i is to assume that the samples are independent as follows for a case where *n*=3 samples within some group *i*:

$$V_i = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

In the above case, the variance is σ^2 and the co-variances (represented by elements off the diagonal) are all zero. To represent sample errors that are dependent within some group$_i$ the variance matrix V_i could be a general correlation matrix

$$V_i = \begin{bmatrix} \sigma^2 & c_{2,1} & c_{3,1} \\ c_{1,2} & \sigma^2 & c_{3,2} \\ c_{1,3} & c_{2,3} & \sigma^2 \end{bmatrix}$$

where there are $\binom{n}{2}$ additional parameters that need to be estimated in addition to the variance. The general correlation matrix is rarely used because the number of parameters that need to be estimated is always larger than the sample size. A useful alternative is the compound symmetric correlation matrix:

$$V_i = \begin{bmatrix} \sigma^2 & \varphi & \varphi \\ \varphi & \sigma^2 & \varphi \\ \varphi & \varphi & \sigma^2 \end{bmatrix}$$

where there is only one other parameter besides the variance that needs to be estimated since any samples in the same group are assumed to have the same co-variance φ.[32] For this dissertation we will use the compound symmetric correlation matrix because with n^2 parameters to be estimated it is impossible to meaningfully fit the general correlation matrix.

1.6 Normalization of 16S counts

In a next-generation sequencing experiment, there are inevitably different numbers of reads that are obtained from different samples. In order to make meaningful statistical comparisons across the samples, throughout this dissertation we will use the following normalization technique [17]:

$$\log_{10}\left(\frac{Raw\ count\ for\ sample\ i}{Number\ of\ sequences\ in\ sample\ i} * Average\ \#of\ sequences\ per\ sample\ i + 1\right)$$

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction to microbiome studies and metagenomics

Bacteria and other microorganisms have proven to be among the most adaptive and diverse of all living things. They are able to live in a range of environments from the soil, aquatic environments and even other living organisms. When other living organisms act as hosts to microbial organisms, the relationship between the host and the microbes can be described as having no effect on the host (communal), being beneficial to host (symbiotic) or being harmful to the host (parasitic). Understanding how the relationship between host and its collective microbes impacts the health of the host is of great importance. This body of work will focus on the techniques used to quantify the composition and functional abundances of microbiota in a host-microbiota system.

2.1.1 Initial microbial community characterization focused on the 16S rRNA gene.

Before the modern techniques of next generation sequencing were developed, identification of the microbial organisms in a community classically involved culture-dependent techniques. A limitation of culture-dependent methods is that they have a tendency to underestimate the diversity of a community since only cultivable species will be observed[33]. An alternative to culturing is the use of culturing-independent methods in which DNA sequences that are conserved across species are used to determine which microbes are present in a community.

The use of molecular sequences to determine phylogeny, or the evolutionary history of organisms, began when Carl Woese and his group found a set of bacteria that are morphologically diverse methane-producing bacteria[34]. Their study sought to discover a way to relate these organisms based on more than physiological traits, since many bacteria look similar but are distant relatives, using the essential highly-conserved 16S small-subunit ribosomal RNA (rRNA) to assign phylogeny. Ten years later, Woese described how the sequences of the 16S rRNA gene can be used to build molecular sequence based phylogenetic trees[35]. 16S rRNA acts as a chronometer because changes in the sequences of the molecule can be used to tell the amount of time since some ancestral sequence. Another factor that makes the 16S rRNA gene an outstanding candidate to be sequenced as a phylogenetic marker is that it has both conserved and variable regions. The conserved regions can be used as hybridization region for primers while the variable regions can be sequenced and compared across organisms to determine phylogeny. There are a total of nine variable regions in the 16S rRNA gene including the regions V1-V3, V3-V4 and V6, which are popular choices for PCR primers[36]. Even to this day, Woese method of target sequencing the 16S rRNA gene is the primary technique used to measure the diversity and abundance of a microbial community.

2.1.2 Shotgun sequencing can yield information about the entire microbial community

Though 16S rRNA gene sequences can give researchers an idea of who is in a microbial community, 16S rRNA gene sequences are not sufficient if one is interested in what the microbes in a community are capable of doing. A culture-free method that can address this problem is whole-genome metagenome shotgun (WGS) sequencing in which all of the DNA in a microbial community[37] is extracted, sequenced and aligned to a database of multiple bacterial genomes. WGS sequencing can yield insights about what the microbes in the community are capable of doing even for those microbes whose whole genome is not yet sequenced.

2.1.3 Using 16S rRNA sequence reads to predict metagenomic data

The cost of WGS sequencing can be quite expensive in comparison to the sequencing of a marker gene such as 16S rRNA[38]. For this reason, many studies only conduct 16S sequencing. However, 16S sequences only give insights about the phylogenetic diversity in a microbial community, but not the functional capability of the microbial community. There have been efforts to predict the functional capabilities of a microbial community using only 16S rRNA sequence reads. One example of a software program that makes these types of predictions is called PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) developed by Curtis Huttenhower and his colleagues[39]. PICRUSt infers metagenomic functional content by taking operational taxonomical units (OTUs), which represent a collection of microbes that share some threshold of 16S sequence similarity, and mapping these OTUs' representative sequence to previously sequenced bacterial genomes. The OTUs that one uses as input for PICRUSt must be generated by mapping sequences to a reference database of 16S sequences. This dissertation will evaluate the use of the PICRUSt algorithm to predict functional abundances.

2.1.4 Metatranscriptomics to understand function of microbial community at a given time

Another important area in the field of microbial genomics is metatranscriptomics in which all the transcripts, RNA sequences generated as a result of transcription, of all microorganisms in a system is collected. Metatranscriptomics is beneficial because one can discover differences in gene expression of a microbial community as a result of a disease state as opposed to the simple metagenomics that can only reveal the genomic capabilities of a community[40, 41].

2.2 Humans and our microbiome

In order to understand the shifts in the makeup and functional capabilities of a microbial community in humans with a given disease, it is first necessary to define microbial community composition in healthy humans. This was the goal of The Human Microbiome Project (HMP)[6, 42] which sought to discover the relationship between humans and the microbes they host in a healthy human cohort of 242 subjects between the ages of 18 and 40. The HMP consortium has shown that across 15-18 sampled body sites, microbes have a distinct distribution of relative abundance and diversity and that there were no taxa common in all of the sites. For example, *Streptococcus* is the prevalent bacterial genus in oral sites, *Lactobacillus* in vaginal sites and *Bacteroides* in the gut. HMP has shown that while the composition of a microbial community, as measured with 16S rRNA sequences, may vary from subject to subject, the functions of genes as measured with whole-genome metagenome sequences are much more consistent across subjects. This variation in 16S sequences from subject to subject occurred even within the same sites providing evidence that although there are common taxa present in certain sites, each individual has their own microbial signature[6]. There have been studies that have shown that not only does the environment of the body site affect the makeup of a bacterial community, but factors such as diet[11, 14], disease[15-17, 43], age[7], and living conditions[20, 22] also play a role in shaping the structure of a bacterial community. However, within the healthy HMP dataset, patient characteristics had little association with the microbiome and that the individual signature in the microbiome is fairly resistant to change over time. Winglee et. al. have argued that while the literature shows that environmental factors such as change in diet can lead to changes in the microbiome, these changes are small compared to the differences in the microbiome between individuals[44]. It has also been shown that in addition to the microbiome of an individual being resistant to change, when the microbiome is perturbed there is a state of equilibrium to which the community tends to return[5].

2.3 Model organisms used for microbiome studies

In order to better understand how our microbiome works it is common to use model organisms to address questions regarding the human gut microbiome. In general, human studies can be challenging because clinicians cannot ensure continued participation outside of incentive, controlled diets are hard to enforce for ethical reasons and the microbiome cannot be directly manipulated without potential harmful consequences to patients. Human studies can also be difficult because of the variation in genotypes and in bacterial species within each individual[45]. Model organisms allow the researcher to control perturbations in the host-microbiota system in a manner that is not feasible in a human-microbiota system. In order to determine which model organism is most appropriate, one must understand what aspects of the host-microbiota relationship the researcher would like to investigate[46]. For example, zebrafish is a simple vertebrate model with a complex microbiota and an adaptive immune system that recognizes the microbes that it hosts. Also, since zebrafish embryos are transparent, their microbes could be fluorescently labeled and visualized in real-time[47]. Zebrafish has been used in a range of gut microbiome research from inflammatory bowel disease[48] to fatty acid absorption[3]. Rodents, however, are the most widely used model organism to study gut microbiome in mammalian systems[45]. This is partly due to the fact that rodents and humans share 99% of their genes[49] and the same two phyla, Firmicutes and Bacteroidetes, dominate the guts of both mice and humans[18] making rodents an appropriate system to study the role that host genetics have on the microbiome[46]. Using rodents as a model host system comes with the benefit of controlling the environment of the experiment. Despite the similarities between human and rodents, a disadvantage to using rodents is that rodents have subtle differences in the structure of the mouth, oropharynx and gastrointestinal tract that could shape differences in microbial communities between rodents and humans[46]. Rodents are also very coprophagic in that they consume their feces. This can complicate any studies that examine gut microbiota.

In general, the microbiome among different mammalian host species can vary vastly. It has been shown that a possible reason for these differences could be because of the differences in the diets of the host species[50]. Ley et al. found that host species who were carnivores and herbivores had distinct microbial communities with omnivores showing a microbial community composition between the two[50]. The number of phylum present in the hosts increases from carnivores to omnivores to herbivores. Ley at al argued that this was because adaptation to a plant-based diet required the digestion of complex carbohydrates leading to the need of more different microbiota to assist with digestion. Therefore, the use of a model organism whose diet is similar to humans, such as an omnivorous non-human primate, may produce studies with insights that cannot be obtained only by use of rodent model systems. Because each model system has distinct advantages, this dissertation will focus on both rodents and primates as a host system.

2.4 Systematic biases in quantifying a microbial community

Microbiome research can be challenging as artifacts caused by systematic biases that are introduced in sample collection, sequencing and data analysis can produce features of the dataset that do not reflect the structure of the underlying bacterial communities. These biases include PCR amplification biases, sampling and sequencing depth biases and the source of the samples [51-55].

2.4.1 PCR amplification bias

In order to produce enough copies of the 16S rRNA gene sequence fragments to generate libraries for sequencing, polymerase chain reaction (PCR) is essential. In spite of the usefulness of PCR amplification, PCR amplification can introduce a number of biases[56]. These biases include primer design which affects how the well primer anneals to the DNA[57, 58] and the number of PCR cycles leading to overrepresentation of a community's richness due to chimeric sequences[59]. Chimeric sequences are formed when amplicons are terminated prematurely and during the next cycle of PCR the amplicon anneals to DNA from a different source[60]. Methods of decreasing the biases due to PCR amplification have been introduced and they include using real-time PCR[57] and de-noising the sequences after the PCR amplification[53]. Although these methods attempt to decrease biases, there is no way to completely avoid amplification biases. One possible way to measure PCR bias is to compare 16S and whole-genome shotgun metagenome sequences using programs such as MetaPhlAn, to make taxonomic predictions derived from WGS sequences [61].

2.4.2 Sequencing depth bias

Insufficient sequencing depth can introduce its own biases. These biases can lead to the underestimation of rare operational taxonomic units (OTUs). By definition, a rare OTUs is absent in most samples [62]. OTUs can be absent for biological reasons (the taxa was truly not present in the sample) or for technical reasons (the OTU was not detected because the sequencing depth was not sufficient)[62]. To avoid this sequencing depth bias, normally the variable region of the 16S rRNA gene is deeply sequenced (sequenced hundreds and sometimes thousands of times over the same region) in order to identify low abundance microbiota, but in practice there is no way to determine if for a given sample, more sequencing might have found a missing OTU.

Lagier et al. attempts to address sequencing depth biases by the use of culturomics[63]. They collected three stool samples (two from lean African males and one from an obese French male) and designed 212 culture settings (varied by physiochemical conditions) and incubated the cultures with the stool samples. Antibiotics, bacteriophages and active and passive filtration was applied to the cultures in order to eliminate dominant gut bacteria so that the culture analysis could be selective for rare taxa. They used a type of mass spectrometry (MS) called Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) to identify microbial species from the 3000 colonies they isolated. The use of the MALDI-TOF MS eliminates the need to use Gram-staining and other biochemical tests to distinguish different species of bacteria in a culture[64]. The bacterial community of the three samples were also identified using the culture-independent method of deep sequencing the 16S rRNA gene targeting the V6 region for comparison. Lagier et al. found that 341 species were identified using culturomics and 698 species using pyrosequencing but the two methods shared only 51 species.

Paulson et al. attempted to address the problem of sequencing biases using a more statistical approach[62]. They introduce two methods: one to decrease biases due to uneven sequencing depth and another to decreasing biases due to undersampling. The first method used to address the issue of disproportionate sequencing depth was to normalize the read counts by the cumulative sum up to some given percentile threshold. For example, for a threshold of 75, raw counts would be divided by the cumulative sum of the top 75th percentile of nonzero counts. The second method was to account for low sequencing depths. Paulson et al. introduced what they call a zero-inflated Gaussian (ZIG) distribution mixture model. This model is meant to determine the probability that a zero OTU count was a result of undersampling or absence of the OTU, and it could be used to help researchers determine if it is more likely that they have a sequencing depth problem or that there is a higher probability of an absence of a given OTU.

2.4.3 Source of samples used to quantify microbial community

The prevalence of particular microbes in a sample could be indicative of the environment in which the sample was taken such as the previously mentioned *Bacteroides* in stool. Depending on the source of the sample, some samples are more prone to environmental contamination than others. In order to elucidate the microbiome of the gut, stool samples are commonly used. However, attempts to standardize fecal sample handling[65] are still ongoing[66] and differences in handling stool samples can lead to differences in relative abundance of bacteria in the samples[26, 67].

Huse et al. have compared the use of biopsy samples of the intestinal mucosa versus samples taken by brushing the intestinal mucosa[68]. They were interested in comparing biopsies with intestinal brushing because biopsies only cover a small area leading to a potentially biased distribution. The group took 16 matched biopsy and brush samples from 4 patients with a history of ulcerative colitis (UC) in a longitudinal study. They found a high degree of similarity between the two techniques (with Pearson's R generally > 0.9), although samples taken using the brushing technique had a higher bacteria DNA to human DNA ratio. From this the authors concluded that brushing is the better technique because there is more bacterial DNA, the brushing covers a larger area and it eliminates infection risk in the subject.

Huse et. al. were particularly interested in microbes that are associated with the intestinal epithelium, so for this reason they did not include stool samples in the study since they felt that stool samples missed mucosa-related taxa. While this may be true, most studies of the gut microbiome use stool samples because of the abundance of DNA sequences in the stool samples, which are easily collected while avoiding the intrusive biopsy. For this reason, it would be beneficial to know how stool samples compare with biopsy samples, brush samples or any other sources used to obtain samples that represent the human gut microbiome. In this dissertation, we include a study designed to determine if rectal swabs can be used as an alternative to stool samples. We will examine this question with a dataset for which we have both swab and stool samples from patients and from those samples we extracted whole genome shotgun sequences and 16S rRNA gene sequences.

2.5 Dietary Fructose and its impact on the American digestive system

2.5.1 Fructose have largely been introduced into the American diet in the form of high-fructose corn syrup (HFCS)

The 1970s introduced high-fructose corn syrup (HFCS) as a replacement for sucrose (table sugar) as a sweetener in soft drinks in the United States. Fructose utilization was driven by availability as fructose is derived from corn, which is in abundance in the Midwest region of the United States and cost compared to that of sucrose. HFCS is a liquid sweetener that is made of a combination of fructose and glucose, usually 55% fructose and 45% glucose[69]. Soft-drinks are the primary source of HFCS in our diets but it is also in present in breakfast cereals, jams, and canned drinks[70]. At the molecular level, HFCS differs from sucrose in that HFCS is a mixture of unbounded fructose and glucose monosaccharides while sucrose is made of disaccharides in which glucose and fructose components are joined by a glycosidic bond[71].

2.5.2 Increased consumption of fructose may have contributed to obesity epidemic

HFCS has been controversial because the increase in usage has paralleled the obesity epidemic in America. It has therefore been suggested that HFCS may play a role in the development of obesity. Bocarsly et al. have shown that rats that were given access to HFCS not only gained significantly more weight than their counterparts who were given equal access to sucrose, but also had higher triglyceride levels and more abdominal fat[31]. Bocarsly et al. have shown that although fructose and glucose are present in similar proportions in the blood stream, the two sugars had different effects on weight gain. This difference could be because fructose from HFCS is metabolized at an earlier point than that of sucrose which could result in unregulated creation of carbon molecules that are transformed into fatty acids[72]. Since one of the main roles of gut microbiota is metabolism it would be of interest to determine if there are differences in the microbiota of subjects fed different types of sugar. This dissertation addresses this problem in a rodent animal model.

2.6 Association of gut microbial dysbiosis and colorectal cancer

Colorectal cancer (CRC) is the third deadliest cancer with close to 50,000 people succumbing to the disease in 2015[73] in the United States alone. CRC has a higher incidence in the U.S. compared to other areas in the world, and it has been suggested that this could be due to the heavy intake of animal protein in the Western diet compared to other diets such as the Mediterranean diet[74]. In the 1940s and 1950s CRC was the deadliest cancer, but the advancement of treatments, early detection tests and the reduction of risk factors such as smoking and eating red-meat have helped to increase the 5-year survival rate to 65.4% [73]. There have been studies that have shown that the lack of protective bacteria in the gut has an effect on the development of colorectal cancer[74]. Bacteria such as *Bacteriodes* and *Prevotella* have a significantly higher abundance in patients with CRC[75]. Sanapareddy et al. found that people who had colorectal adenomas, which is a benign tumor that is precursor to CRC in many cases, had an overall higher gut microbial richness than people without colorectal adenomas. Many of the bacteria contributing to the higher microbial richness were pathogenic or belonged to the Proteobacteria phylum[17]. It is not clear whether or not these bacteria that are associated with CRC actually cause CRC. There is still more to be understood about the impact of the gut microbiome on CRC before microbial interventions are introduced as a tool to combat CRC. Despite this incomplete knowledge, knowing which bacteria are more prevalent in CRC patients might make it possible to develop a non-invasive screening tool to detect CRC. Such a screening tool has been proposed by Zackular et al[76]. which they used a Bayesian model to predict whether a person was healthy, had colorectal adenomas or had CRC using microbiota abundances as the input variables. In this dissertation, we determine if the same microbiota that are associated with colorectal adenomas in one dataset are same microbe that are associated with colorectal adenomas in other datasets.

2.7 Conclusion

As previously described humans are hosts to thousands of species of bacteria and the symbiotic relationship between humans and our microbiota can be both commensal and parasitic. To understand these relationships, it is imperative that we acknowledge the many biases and inconsistencies that are introduced when we attempt to quantify the distributions of the organisms and their gene families and attempt to work towards a standardization of methods used to quantify the structure and function of a microbial community. This dissertation aims to do so by calculating the variance of microbiota and gene family classifications measured using different sample sources, different 16S rRNA databases and sequence types and to apply these methods to ask how does sugar intake impact the microbial community and how consistent are changes to the microbial community associated with colorectal adenomas.

CHAPTER 3: COMPARISON OF MICROBIAL COMMUNITIES OF SWAB AND STOOL SAMPLES

3.1 Abstract

3.1.1 Background. Stool samples are the standard sample type from which sequences are extracted in human gut microbiome studies. In spite of this, in clinical trial and citizen science there is a growing interest in using swab samples to represent the gut microbiome because it is easier to collect and handle especially when subjects are collecting the samples themselves.

3.1.2 Methods. Here we use 16S rRNA sequence reads to assign taxonomy in a gut microbial community in swab versus stool samples and we also compare the use of WGS sequence reads to elucidate gene families present in a gut microbiome in swab versus stool samples.

3.1.3 Results. We found that the taxonomic classifications generated by 16S sequences, exhibited a large difference between swab samples and stool samples.

3.1.4 Conclusion. We conclude that swab samples are unlikely to replace stool samples as they generate a different picture of gut microbial community composition.

3.2 Introduction

Microbiota in the human colorectum are responsible for a substantial number of physiological functions within the gut that have both localized and systemic effects including immunity, nutrient metabolism, growth, and energy harvesting[77-80]. Advances in next-generation sequencing promise to yield new insights into the role of the microbiota in health and disease. Compositional shifts in the diversity or relative distributions of members of the gut microbiota have begun to be linked to several diseases including atherosclerosis[81], obesity[82], and inflammatory bowel disease[78, 83, 84] among others.

The function and prevalence of microbiota within the colorectum likely varies by niche (i.e. luminal vs. adherent mucosa)[85-87]. To elucidate the true composition of the microbiota of the total colorectum, it is therefore necessary to sample various site in the colorectum via biopsies. Unfortunately, collection of mucosal samples by biopsy is a highly invasive procedure with risks of perforation that make it unfeasible for large scale studies. Further, studies have found the colon has one of the steepest oxygen gradients in the body, which reduces rapidly from the mucosa to near anoxia at the middle of lumen [88]. Thus, luminal microbes are more likely to be anaerobic than mucosal communities [85]. Anaerobic bacteria may play a key role in fermentation and metabolism of luminal contents (e.g. nutrients or carcinogens)[88-90] while mucosal bacteria may possibly be involved with autoimmune functions. In addition, previous studies have found adherent mucosal communities are less diverse than luminal bacteria although they share many of the same predominant species[86-89, 91, 92]. Rectal swabs may prove to be a simple and inexpensive collection method that samples mucosal communities. Comparisons between swab and mucosal biopsy samples have found swab samples may be capable of capturing many of the same bacteria as mucosal biopsy samples but also may be different than fecal samples[27, 85, 91].

Differences in microbial communities are not only driven by niche, but are also by the individual that hosts the microbial community. There have been many studies, including the Human Microbiome Project[6, 66], that have observed large individual differences in the microbiome leading to the conclusion that each person has an average unique microbial signature[4, 87, 93, 94]. These studies have used fecal or biopsy samples, however, and there have been no studies showing the stability of the microbiome for rectal swab samples. It has been proposed that individuals can be grouped into three "enterotypes"[95], however, the three enterotypes have been shown not to remain constant across time[10, 94, 96] and the enterotype abstraction has proven controversial[96-98]. Furthermore, since the enterotype studies were all based on fecal samples, the identified enterotypes may be highly weighted by luminal anaerobic microbiota and may not represent well the entire microbiome of the colorectum[88].

Strong individual signatures over time have also been evident in longer longitudinal studies. Rajilić-Stojanović et al. collected fecal samples up to 9 times from 5 individuals over a decade and found that although there were some changes in abundances with age, individual-specific patterns persisted[94]. The findings from their study also indicated a single spot fecal sample was not able to capture the presence of all core colonizers. Thus, in order to plan the most effective large-scale studies, it is necessary to evaluate whether multiple collections at different times may be more informative than one spot sample to define an individual's microbial signature and whether rectal swab alone or in combination with fecal samples yields a significantly better representation of the gut microbiota than fecal collection alone.

Based on these needs, in this study, we first compared the microbial composition of rectal swabs versus fecal samples. We then addressed the question of whether or not a microbial signature could be as strongly detected in swab samples as has been previously demonstrated in fecal samples. Lastly, we compared the functional content of swab versus stool samples whole-genome shotgun (WGS) sequencing. We hypothesized that rectal swabs have a different bacterial composition than fecal samples but are not so different in functional composition. We also found that two collection time points may more reliably reflect the long-term diversity of microbiota than a single spot collection.

3.3 Methodology

3.3.1 Study Population

The participants in this study were selected from the Personalized Prevention of Colorectal Cancer Trial (PPCCT), which is an on-going, double-blind, placebo-controlled, randomized clinical trial of 12 weeks of personalized magnesium supplementation designed to test magnesium and the interaction between *TRPM7* genotype and reduction of calcium/magnesium intake ratio by magnesium supplementation on colorectal carcinogenesis biomarkers. From the parent study for sample selection for this analysis, eligible participants were 40-85 years of age, in good health, able to participate in a low-to-moderate intensity supplement intervention, had a personal history of colorectal hyperplastic and/or adenomatous polyps, had known *TRPM7* rs8042919 genotype, and, based on two 24-hour dietary recalls, had daily calcium intake between 700-2000 mg/day and a ratio of daily intake of calcium and magnesium greater than 2.6. Participants were identified from two on-going studies of colorectal polyps[99], or from medical record review of individuals diagnosed with colorectal polys at Vanderbilt University between 4/14/1995 and 11/22/2013. Exclusion criteria included any personal history of cancer other than non-melanoma skin cancer, colon resection or colectomy, gastric bypass, organ transplantation, inflammatory bowel disease, chronic diarrhea, chronic renal diseases, hepatic cirrhosis, chronic ischemic heart disease, or Type I diabetes mellitus. Also excluded were individuals using medications that may potentially interact with magnesium, or who were breastfeeding or pregnant. Eligible participants were randomized to receive either placebo (microcrystalline cellulose) or magnesium supplementation (combinations of pills containing 104.1 mg, 77.25 mg, and 73.35 mg elemental magnesium as glycinate) for twelve weeks. Participants, health care providers and investigators were blinded to treatment assignments. The study was approved by the Vanderbilt Institutional Review Board. Participants included in this analysis were selected and assayed at two different time points from individuals who had completed the trial at the time of selection and who provided relevant samples at the beginning and end of the trial period. Participants were excluded from selection if they used oral or injected antibiotics in the past 12 months before the study or during the study period. From these criteria, individuals were randomly selected such that 50% were from the placebo arm and 50% from the treatment arm. A total of 60 individuals were selected from 150 participants enrolled between 4/11/2011 and 12/11/2013.

3.3.2 Sample Collection

Every participant was asked to collect a stool sample at home up to three days prior to their three in-person clinic visits. Participants were provided with instructions and a kit to collect four sterile vials of stool from a single bowel movement. The samples were immediately frozen in their home freezer. They were also provided with a Styrofoam cooler and ice pack to use to transport the sample to their visit where they gave their sample to the research staff. Upon receipt, the samples were placed in -80˚C freezers for future analysis.

At the baseline and final clinic visits, the study physician inserted a culturette swab about 2 inches into the rectum, swabbed the rectal mucosa, and immediately placed the swab into the storage vial. Then the physician took the biopsies at 10 cm above the anal verge through an anoscopy and put the fresh biopsy specimens into separate storage vials. All the samples were immediately frozen at 80˚C until use. No colon cleansing preparation was used prior to collection. For the participants selected for this study, the mean (standard deviation) days between the first sample collection (clinic visit 1) and the last sample collection (clinic visit 3) was 86.4 (6.6) days.

3.3.3 Microbial Classification and Assignment of Functional Content

3.3.3.1 Preprocessing of 16S Sequences

Raw Illumina base call outputs (BCL) obtained from the MiSeq were converted, but not demultiplexed, to paired-end fastq files using CASAVA[100]. The resulting paired-end fastq files were joined into single-end reads using fastq-join[101, 102]. Quality filtering was applied to the output of fastq-join requiring that greater than 80% of the base pairs be specified with a quality score of at least 25 in order for a read to be retained. The quality filtered reads were then demultiplexed. Reads whose index sequence was not an exact match to the specified barcode were eliminated (Table 3.1). Demultiplexed reads followed the QIIME[103] split_libraries.py output convention and were suitable for subsequent analysis.

3.3.3.2 Phylogenetic Assignment of Reads

Demultiplexed reads were passed through QA/QC filtering pipeline leaving 73,665,484 sequences reads that were then classified using a naïve Bayesian classifier (version 2.10.1 of the RDP classifier[104, 105]). Additionally, the 16S rRNA gene sequence reads were clustered into de novo Operational Taxonomic Units (OTUs) with 97% identity using QIIME. OTU abundances were normalized using the normalization technique described in Section 1.6[17].

3.3.3.3 Whole Genome Shotgun (WGS) Functional Classification

Whole-genome shotgun metagenomics DNA sequencing was conducted for 50 participants (42 during the first wave of selection and assay and 8 during the second wave) including 100 stool samples, 28 rectal swabs, and 16 tissue samples. Paired-end FASTQ files containing WGS sequences were converted to FASTA format. After filtering sequences mapping to the human genome forward WGS sequence reads were then queried against the KEGG gene family protein database[106, 107] using BLAST[108]. Only BLAST hits that had an E-value equal to or less than 1×10^{-3} were kept. KEGG pathway abundances were calculated using HUMAnN[109]. KEGG gene families and pathways which were not present in at least 20% of the samples were removed (Table 3.2) (Figure 3.1). The counts were then log-normalized (Section 1.6).

3.3.3.4 Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) was performed on the data generated by RDP classifier (microbial classification) and BLAST (functional classifications) using Bray-Curtis dissimilarity. The R package “vegan”[110] was used to calculate the MDS axis.

3.3.4 Statistical Analysis

Descriptive statistics of mean (standard deviation) for continuous variables and frequencies for categorical variables were derived for characteristics of the study participants. For the analysis of 16S data, the R package “lme4” was utilized to perform a mixed linear model to evaluate the amount of variance due to stool vs. swab measures:

$$MDS\_axis\ OR\ taxon = sampleType + timepoint + treatment + (1|participant) + \ e, \text{ \{Model 1\}}$$

In this model, the type of sample (stool or rectal swab), whether or not the sample was from a participant given a magnesium treatment and time are fixed effects while participant (ID of participant) is a random effect. The model was evaluated for the first 15 MDS axes and all taxa. ANOVA was used on the mixed models to test the null hypothesis that sample type (stool versus swab), magnesium treatment and time did not contribute to the model.

In order to determine if there was a difference in functional composition between the stool and swab sample types, two types of statistical tests were performed. First, mixed linear models were evaluated using matched stool and swab samples (n=14 participants) with participant as a random variable. Second, mixed linear models were performed using distinct participants for stool and swab samples. There were two samples from each participant representing two different time points.

$$MDS\_axis\ OR\ KEGG\_function = sampleType + timepoint + treatment + (1|participant) + \ e, \text{ \{Model 2\}}$$

We performed each of these test for all MDS axes, KEGG pathways and gene families.

3.4. Results

3.4.1 Comparison of Taxa from 16S rRNA Amplicon Sequencing between Stool and Rectal Swab Samples

In order to compare stool and rectal swab sampling of the gut microbiome, we sampled 60 patients from our on-going clinical trial on the effect of magnesium supplementation on development of colorectal cancer (see methods). In an initial analysis, samples were subjected to 16S rRNA sequencing. We found substantial areas of similarity and differences between stool and swab samples based on the MDS ordination of 16S rRNA amplicon sequencing data (Figure 3.2). The first two MDS axes, which accounted for nearly 27% of the total variation of the microbial communities in the samples, showed almost entirely distinct clusters between the stool and swab. However, the 4th MDS axis, which accounts for 5.1% variance of our data, shows almost no separation between stool and swab but instead shows strong clustering based on the subject ID.

To evaluate the statistical significance of these differences in MDS axes, we generated a mixed linear model with time point and sample origin as fixed terms and subject as a random term (Model 1; Table 3.3). We also ran the model with participant as a fixed variable and we have found similar results (data not shown). For the analysis of the first 15 MDS axes, the first few MDS axes were significantly associated with stool versus swab and also associated with participant. However, there was little evidence of separation that is associated with the time point in which the sample was collected (Figure 3.3). Taken together, our modeling suggests that some taxa are highly sensitive to a certain collection type (stool versus swab) while other taxa can be detected with both methods. With both methods, however, there is a strong tendency towards stability of each individual microbiome as changes with time are not pronounced.

To explore which taxa are more sensitive to sampling method, we built mixed linear models for the log-normalized relative abundance of individual taxa at the phyla (Table 3.3) and family levels (Figure 3.4). At the phylum level, 7 out of 8 phyla were significantly different between stool and swab samples, all phyla were significantly associated with participant but only 3 of the 8 phyla was associated with the time in which the sample was taken. We observe substantial taxa-by-taxa variation in that there are particular families that vary by sample type but not participant (e.g. Thermaceae), participant but not sample type (e.g. Desulfovibrionaceae), or 3) both participant and sample type (e.g. Enterobacteriaceae) (Figure 3.5).

Using a chi-square test of independence, we found that there was a significantly ($p < 0.0219$) higher proportion of aerobic genera that were significantly more abundant in swab than in stool including *Acinetobacter*, *Anoxybacillus* and *Geobacillus* (Appendix A). This is in line with our hypothesis that there may be a decreasing aerobic microbiota gradient radially inward towards the lumen of the colorectum.

3.4.2 Comparison of WGS Functional Classifications between Stool and Rectal Swab

In addition to classifying the bacteria composition of swab and stool samples, we also wanted to determine if there is a difference in the functional capabilities of the two types of samples. To address this, we have collected WGS sequences. As was the case for 16S sequences, there are clear differences between stool and rectal swab based on MDS ordination of WGS sequences (Figure 3.6). In order to statistically evaluate these differences, we consider statistical models using two sets of samples: the first model was built with only samples for which we had both swab and stool samples for the same subjects and evaluated as a paired statistic with n = 28 pairs; the second model was built comparing 70 samples taken from stool and 28 samples taken from swab where there were no overlapping samples from the same subjects. (see Methods for details). These tests were performed for all MDS coordinates (Table 3.4) (Figure 3.7), KEGG gene pathways (Table 3.5) (Figure 3.8) and KEGG gene families (Figure 3.9). We found that a substantial number of functional families and pathways were significantly different among stool and swab samples under either statistical model.

The pathways in which we observe significant differences by sample type including those that were related to the KEGG pathway group classified as “human diseases” and were higher in swab samples than in stool samples (Figure 3.8). While there are many KEGG gene families and gene pathways that are associated with sample type, we also observe the abundances of KEGG gene pathway category “Genetic Information Processing”, which includes the KEGG pathways “transcription” and “translation”, are associated with participants.

3.5 Discussion

The goal of this chapter was to determine differences in the microbial communities of rectal swab samples versus that of stool samples taken from 60 participants who were a part of the Personalized Prevention of Colorectal Cancer Trial (PPCCT) and had a history of colorectal adenomas. From these participants we obtained samples from two time points from both rectal swabs and stool from which we performed 16S rRNA sequencing. Using these 16S sequence reads we found that there were indeed differences in the makeup between stool and swab samples. This difference in the structure of the gut microbiome appears to be driven by the increase of anaerobic microbes that are higher in abundance in rectal swab samples than in stool samples. Despite the structural differences due to sample type, the individual signature of each participant was still strongly evident and stable over time.

In addition to 16S rRNA gene sequences, from a subset of the participants we extracted whole-genome shotgun (WGS) sequences from rectal swab and stool samples. Using these WGS sequence reads we aimed to elucidate differences in the microbial gene pathways of the microbial communities between the two sample types. The microbial gene pathways that are significantly different between the two sample types are those pathways belonging to the groups “human disease” and “organismal systems”. Most of these significantly different pathways have a higher abundance in swab samples than in stool samples. We also see that there are microbial gene pathways that are associated with participants and these gene pathways belong to the immune system pathway and pathways relating to genetic information processing such as translation and transcription. This shows us that despite the differences between WGS sequences of the two sample types we are still able to observe the participant’s microbial signature in the microbial gene pathways.

Overall our work suggests that when designing studies that aim to investigate the gut microbiome, the sample that is best to use for sequences extraction highly depends upon what about the colorectum one wishes to understand. For example, in obesity studies in which it is most appropriate to study the bacteria that are involved with metabolism[14, 18, 19, 80] it may be best to use fecal samples since those samples have a higher abundance of anaerobic bacteria which is more dominant in the center of the lumen. On the other hand, if the association of microbes and tumor progression is the aim of a study, then it might be best to collect swab samples from the patients since these samples are likely to pick up microbes that are closer to the epithelium in which is when a tumor would begin to grow.

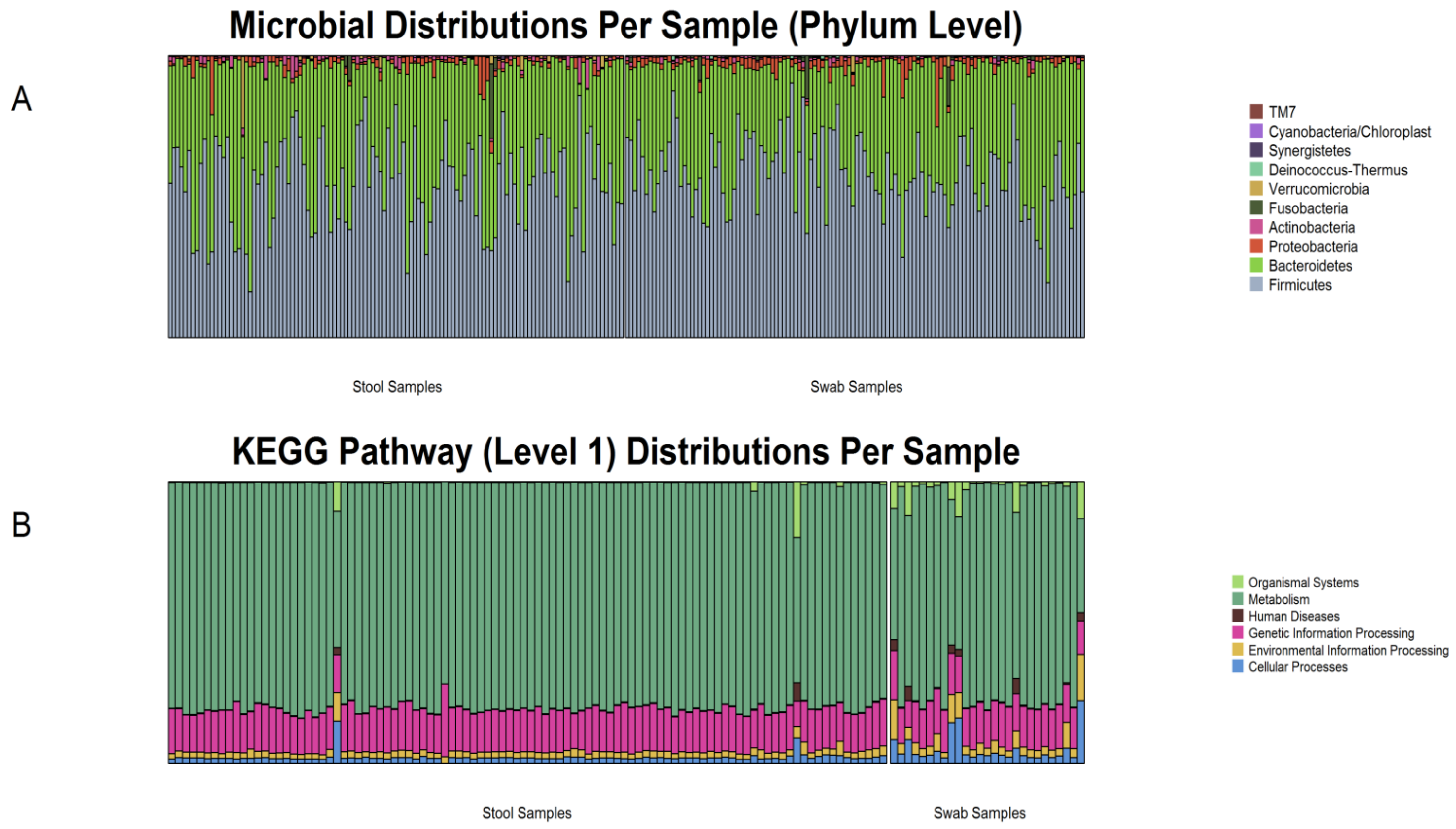

Figure 3.1: Distribution of RDP calls at the phyla level (A) and distribution of KEGG gene pathways at the highest level (B). The distributions of RDP calls varies greatly between samples. However, the distributions of KEGG pathway distributions for stool samples have little variation for the most part.

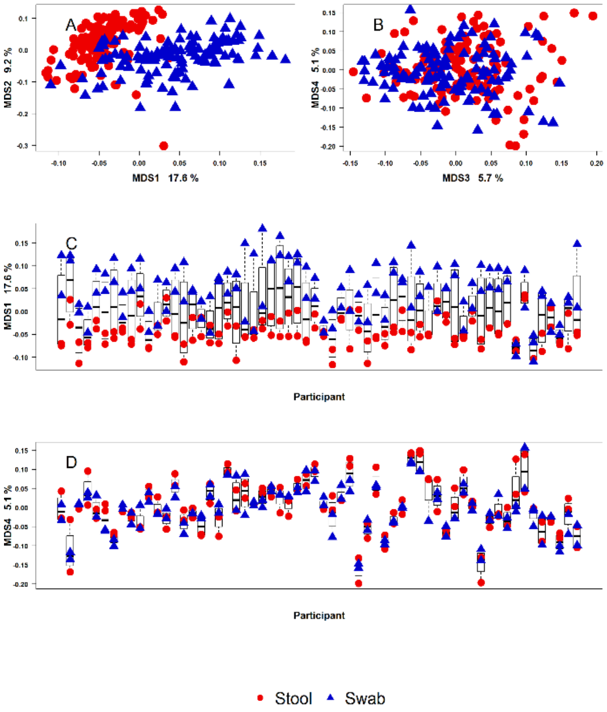

Figure 3.2: Multidimensional scaling (MDS) of RDP calls at the family level. There are four samples (one "pre" and one "post" treatment for both stool and swab) from each of the 60 participants in our study colored by sample origin. The distinct separation of colors shows that there is separation by sample type in MDS axis 1 and MDS axis 2 (A,C) but not in MDS axes 3 and 4 (B). However, the variation in the mean of the bar plots proves clustering by participant in MDS axis 4 (D).

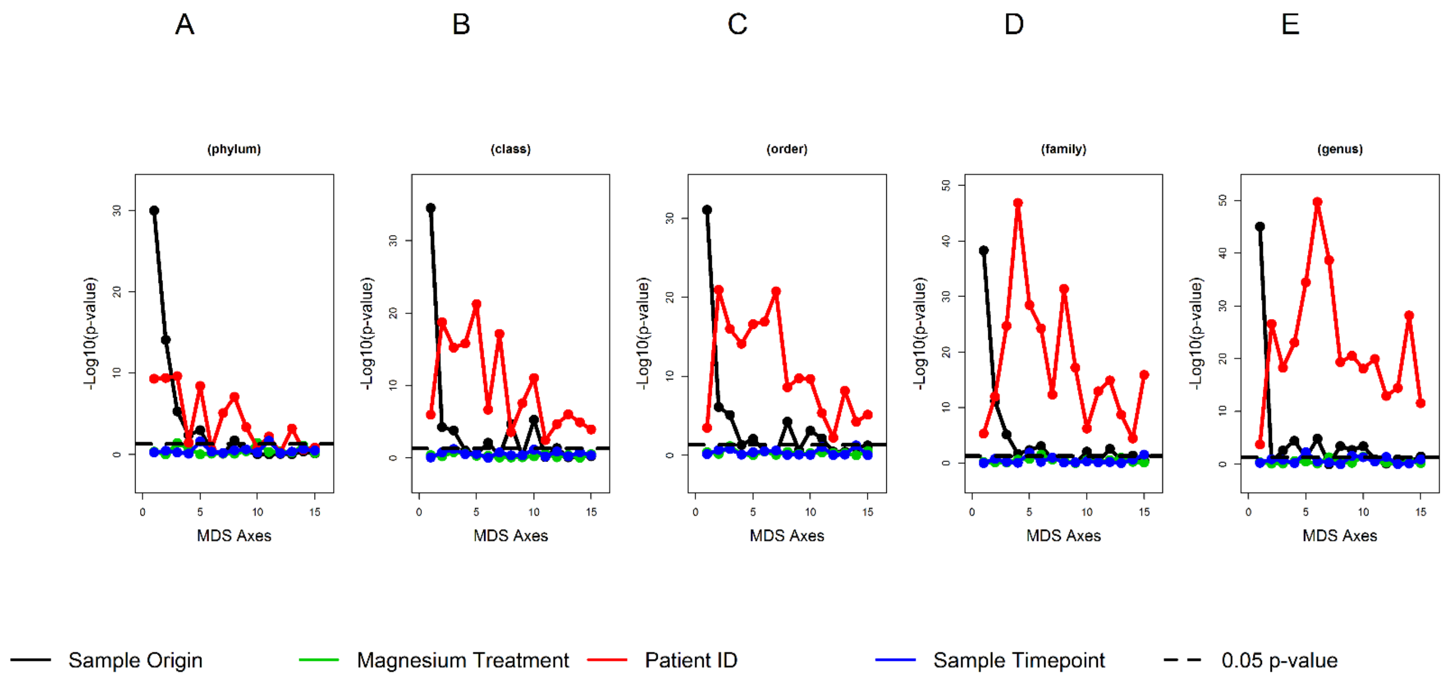

Figure 3.3: First two MDS axes shows strong separation based on the origin of the sample (stool or swab origin), however subsequent MDS axes proves that samples from the same patient tended to exhibit a microbial signature that became more pronounced the higher the taxonomy level. Each of the first 15 MDS axes were dependent variables for the model described by Equation 3. The *p*-values for each independent variable in the model are plotted. We also performed an ANOVA test to test the significance of patient ID in Equation 2 versus Equation 3. The *p*-values for the ANOVA test are shown by the red line. This was performed at the phylum, class, order, family and genus levels.

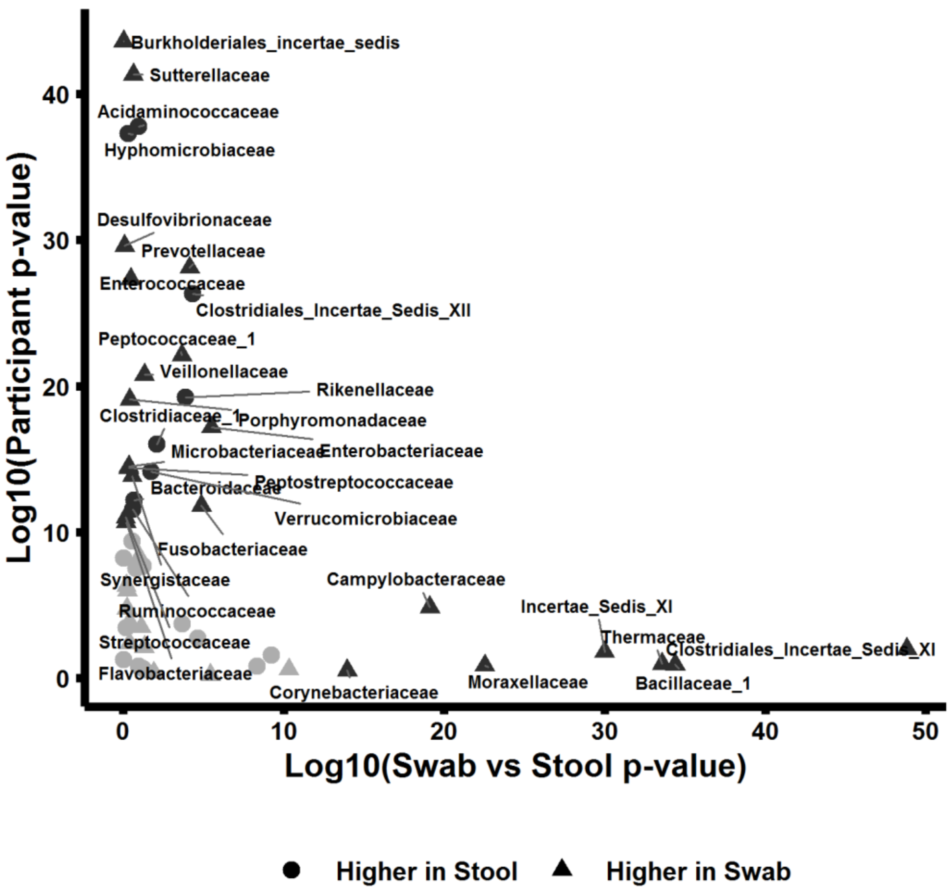

Figure 3.4. Some bacteria families are significantly different between sample type while others differ by study participant. For each taxa at the family level present in at least 25% of samples, *p*-values for a null hypothesis of no difference by stool vs. swab vs. by participant.

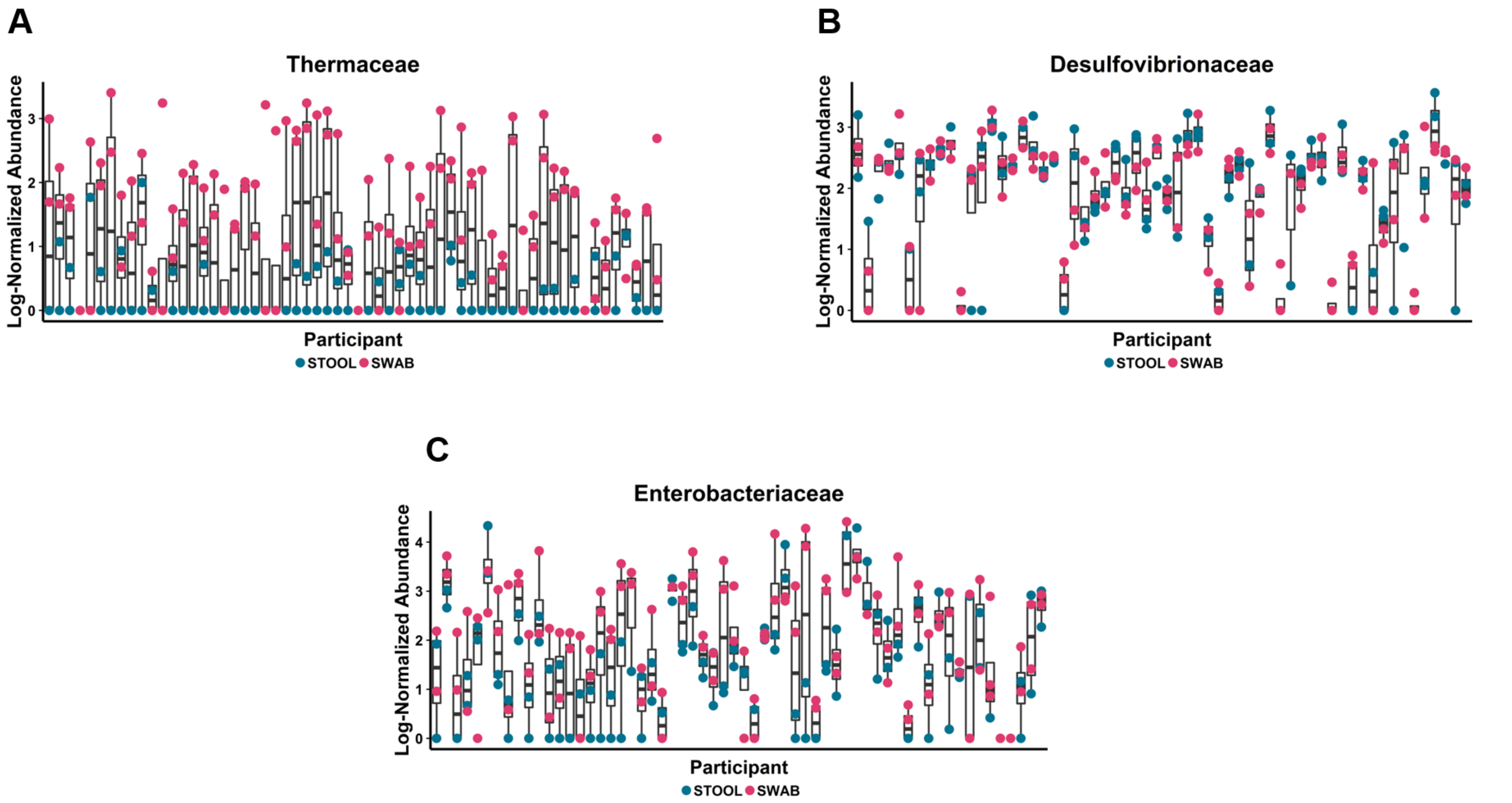

Figure 3.5. Example taxa for which the effect on taxa variation is mostly affected by A. sample type, B. participant or C. both. Points in red represent samples from stool while those in blue are ones from swab.

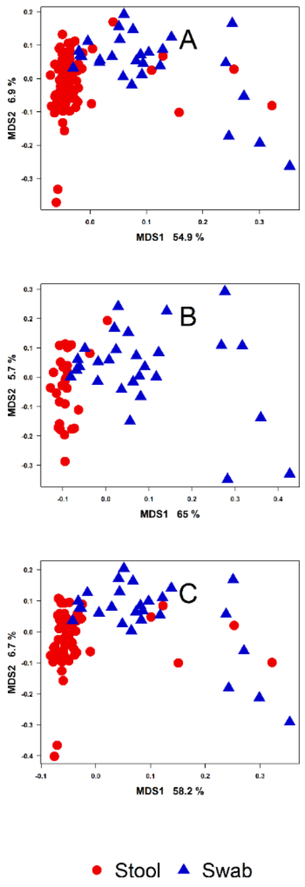

Figure 3.6. Plot of first two coordinates of an MDS ordination of the KEGG gene pathways (level 3) abundance table for WGS swab samples, and stool samples. A. All samples B. samples from the same participant and C. samples from distinct participants

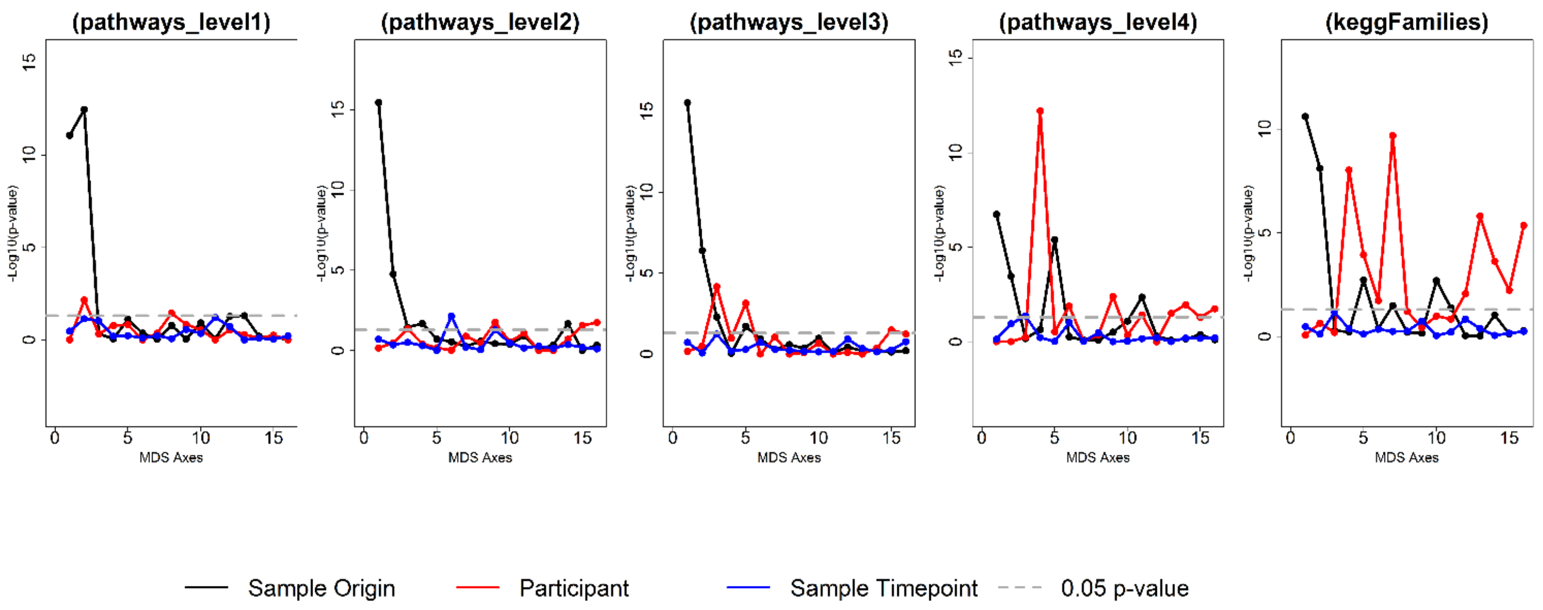

Figure 3.7. The first 15 MDS axes generated from microbial gene pathways were regressed against sample origin (swab versus stool), participant and time point. We see that while there are significant differences in the first MDS axis in stool vs swab vs tissue samples (or only stool vs swab samples in bottom panel), the MDS axes thereafter are significantly different between the participants.

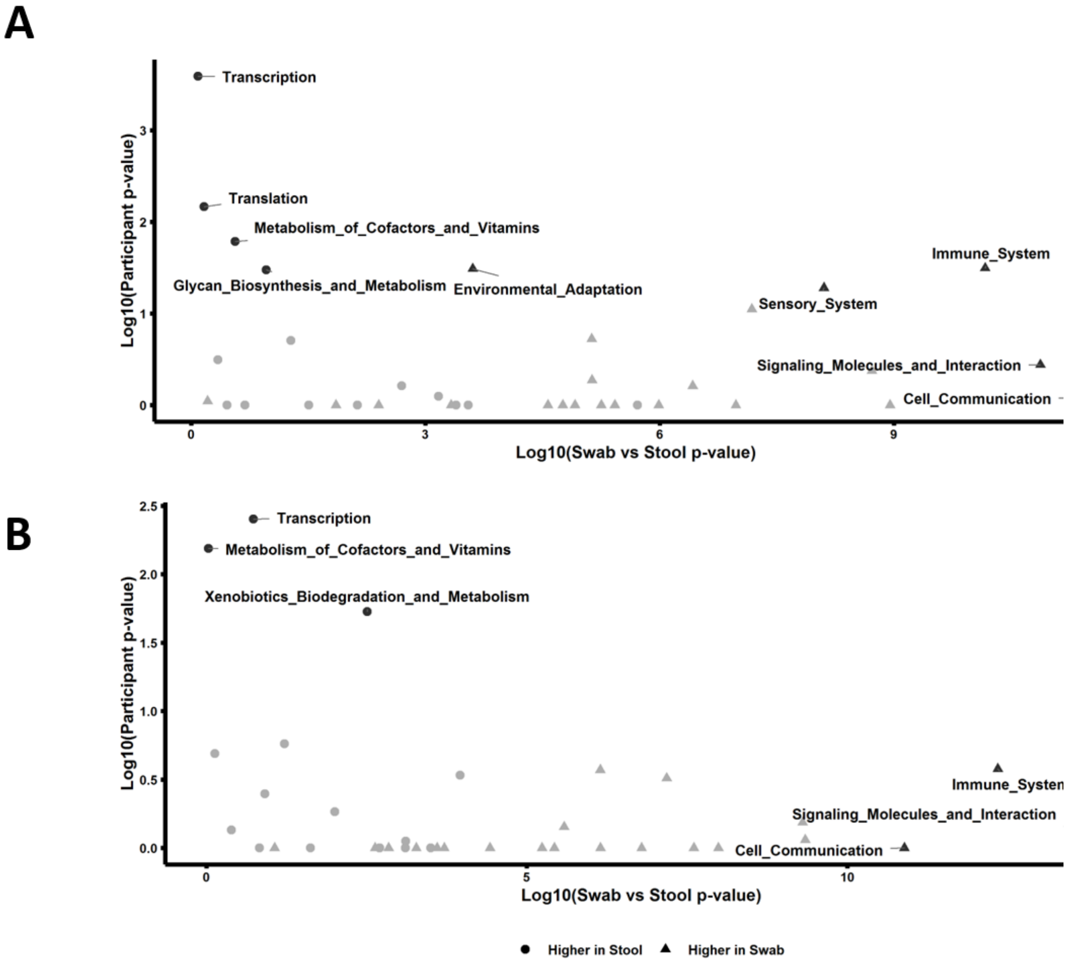

Figure 3.8: Abundances of some KEGG pathways differ by participant while others differ by sample type.

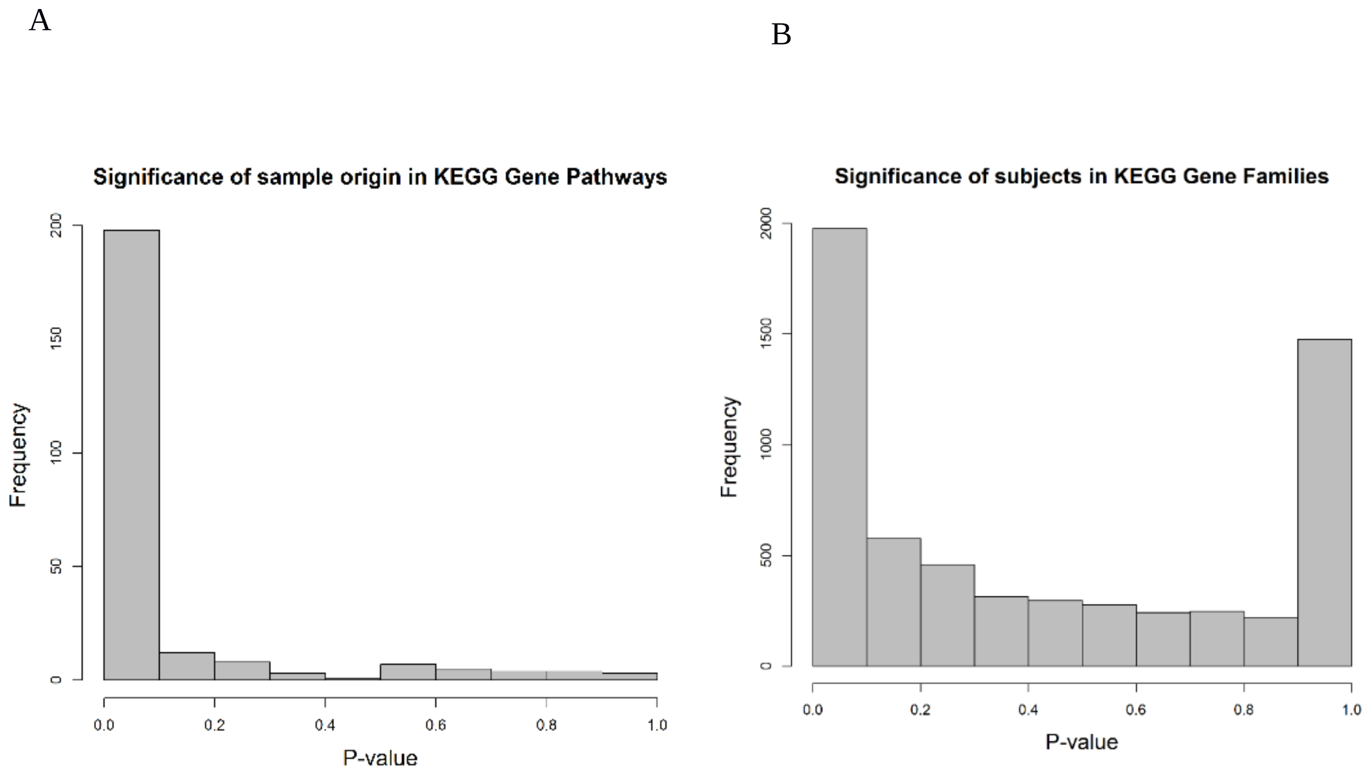

Figure 3.9: The distribution of *p*-values shows that there is a strong separation by sample origin and a fair amount are associated with participant ID. We regressed each of the KEGG gene families against sample origin with participants' ID as a random variable. We plotted the distributions of the *p*-values of the sample origin (A) and participant ID (B).

Table 3.1: Statistics of 16S rRNA sequence reads from both stool and swab samples after various filtering steps.

	Number of Samples	**Number of OTUs**	**Total Number of Sequence Reads**	**Mean Reads per sample ± SD (SE)**	**Minimum reads per sample**	**Maximum reads per sample**
16S reads generated	240		36,832,742	153,469.76 ± 214,793.55 (13,864.86)	85	1,768,150
After clustering into OTUs	240	8,375	36,826,591	153,444.13 ± 213,926.13	85	1,768,150
After filtering out OTUs in less than 20% of samples	240	1,849	32,222,426	134,260.11± 209,547.62	85	1,768,150

Table 3.2: Statistics of whole-genome metagenome shotgun sequence reads from both stool and swab samples after various filtering steps.

	Number of Samples	Total Number of Sequence Reads	Mean Reads per sample ± SD (SE)	Minimum reads per sample	Maximum reads per sample
After removing reads mapping to the human genome	128	25,720,978	200,945.1 ± 104,515.8 (9,237.98)	342	521,089
After removing samples with low read counts	127	25,720,636	202,524.7 ± 103,384.5 (9173.9)	13,405	521,089
After assigning reads to gene families	127	22,285,207	175,474.1 ± 97,062.82 (8612.93)	13,863	455,173

Table 3.3: Differences in 16S sequences due to sample source (stool vs. swab) and participant source. [a] *p*-value derived from mixed model in which participant is random term and sample origin and time point are fixed terms. The *p*-values shown are adjusted at Benjamini Hochberg correction threshold of 10% [b] Limited to phyla present in at least 25% of samples.

Item		**Stool vs. Swab *p*-value[a]**	**Participant *p*-value[a]**	**Time point *p*-value[a]**
Analysis of MDS axes (phylum level)				
Axis	**Percentage Explained**			
MDS1	18.97	$1.43x10^{-30}$	$4.65x10^{-10}$	0.562
MDS2	16.41	$1.58x10^{-15}$	$4.03x10^{-10}$	0.360
MDS3	11.49	$9.25x10^{-06}$	$3.72x10^{-09}$	0.519
MDS4	8.81	0.004	0.055	0.742
MDS5	7.39	0.001	$2.86x10^{-09}$	0.024
MDS6	5.21	0.352	0.217	0.415
MDS7	4.58	0.771	$6.41x10^{-06}$	0.625
MDS8	3.04	0.018	$6.49x10^{-08}$	0.308
MDS9	2.56	0.385	0.0009	0.277
MDS10	2.30	0.906	0.230	0.565
MDS11	2.13	0.969	0.006	0.022
MDS12	1.81	0.813	0.500	0.734
MDS13	1.57	0.907	0.0006	0.437
MDS14	1.26	0.419	0.346	0.236
MDS15	1.08	0.752	0.139	0.298
Analyses of Phyla[b]				
Actinobacteria		0.0005	0.0002	0.330
Bacteroidetes		0.353	0.001	0.048
Deinococcus.Thermus		$3.54x10^{-34}$	0.054	0.086
Firmicutes		0.013	$3.74x10^{-08}$	0.072
Fusobacteria		$3.11x10^{-09}$	$5.49x10^{-13}$	0.452
Proteobacteria		$7.43x10^{-05}$	$1.58x10^{-07}$	0.440
Synergistetes		0.109	$6.81x10^{-15}$	0.386
Verrucomicrobia		0.004	$8.31x10^{-15}$	0.345

Table 3.4: Differences in KEGG gene pathways (level 2) inferred from WGS sequences due to sample source (stool vs. swab) and participant source (distinct participants). [a] *p*-value derived from mixed model in which participant is random term and sample origin and time point are fixed terms. The *p*-values shown are adjusted at Benjamini Hochberg correction threshold of 10%

Item		**Stool vs. Swab *p*-value[a]**	**Participant *p*-value[a]**	**Time point *p*-value[a]**
Analysis of MDS axes				
Axis	**Percentage Explained**			
MDS1	72.9	2.04×10^{-14}	0.313	0.195
MDS2	5.84	1.18×10^{-05}	0.158	0.416
MDS3	2.70	0.040	0.019	0.291
MDS4	2.06	0.026	0.170	0.483
MDS5	2.01	0.188	0.306	0.963
MDS6	1.37	0.291	0.432	0.007
MDS7	1.24	0.500	0.055	0.622
MDS8	1.05	0.309	0.222	0.869
MDS9	1.04	0.383	0.006	0.030
MDS10	0.95	0.964	0.121	0.293
MDS11	0.90	0.114	0.031	0.654
MDS12	0.78	0.883	0.478	0.542
MDS13	0.63	0.463	0.500	0.711
MDS14	0.51	0.013	0.103	0.380
MDS15	0.50	0.958	0.011	0.586

Table 3.5: Differences in KEGG gene pathways (level 2) inferred from WGS sequences due to sample source (stool vs. swab) and participant source (distinct participants).

Item	**Stool vs. Swab p value**[a]	**Participant p value (BH adjusted)**	**Time point p value (BH adjusted)**
Metabolism of Terpenoids and Polyketides	0.068	0.732	0.804
Metabolism of Cofactors and Vitamins	0.317	0.202	0.569
Transcription	0.817	0.010	0.823
Signaling Molecules and Interaction	2.47x10^{-10}	0.999	0.840
Nucleotide Metabolism	0.001	0.999	0.840
Immune System	8.39x10^{-10}	0.206	0.980
Xenobiotics Biodegradation and Metabolism	0.497	0.999	0.840
Membrane Transport	0.390	0.999	0.840
Cell Growth and Death	0.652	0.999	0.840
Folding Sorting and Degradation	0.006	0.999	0.677
Cancers	2.81x10^{-05}	0.999	0.843
Circulatory System	1.59x10^{-05}	0.999	0.597
Transport and Catabolism	1.04x10^{-08}	0.999	0.804
Development	1.55x10^{-06}	0.999	0.569
Excretory System	5.56x10^{-05}	0.999	0.840
Neurodegenerative Diseases	4.90x10^{-07}	0.999	0.804
Metabolism of Other Amino Acids	0.041	0.999	0.569
Glycan Biosynthesis and Metabolism	0.135	0.206	0.804
Nervous System	0.019	0.999	0.569
Lipid Metabolism	0.245	0.999	0.981
Cell Motility	3.50x10^{-07}	0.415	0.569
Infectious Diseases	1.16x10^{-05}	0.999	0.804
Translation	0.706	0.126	0.840
Signal Transduction	3.79x10^{-05}	0.999	0.569
Immune System Diseases	1.83x10^{-05}	0.732	0.840
Replication and Repair	0.003	0.999	0.843
Carbohydrate Metabolism	0.0005	0.999	0.843
Cardiovascular Diseases	1.42x10^{-08}	0.999	0.823
Environmental Adaptation	0.0005	0.206	0.840
Sensory System	4.86x10^{-08}	0.278	0.979
Biosynthesis of Other Secondary Metabolites	0.011	0.999	0.840
Amino Acid Metabolism	6.46x10^{-06}	0.999	0.823
Endocrine System	0.0008	0.999	0.843
Digestive System	1.83x10^{-05}	0.999	0.843
Cell Communication	1.84x10^{-10}	0.999	0.843
Energy Metabolism	0.0007	0.999	0.843
Metabolic Diseases	3.81x10^{-06}	0.999	0.968

CHAPTER 4: EVALUATION OF PREDICTED MICROBIAL GENE PATHWAYS IN A NON-HUMAN PRIMATE SYSTEM

4.1 Abstract

4.1.1 Background. Whole-genome shotgun (WGS) sequencing is expensive in comparison to the target sequencing of a marker gene such as 16S rRNA. 16S sequencing, however, only provides insights about the phylogenetic diversity and not the functional capability of the microbial community. Here we evaluate an algorithm that attempts to predict the functional capabilities of a microbial community using only 16S rRNA sequence reads.

4.1.2 Methods. Our study involved vervets (*Chlorocebus aethiops)* that were either fed a high fructose or control diet. The vervets were housed in cages with one monkey per diet in each cage. Stool was collected over the course of the study and both 16S rRNA gene sequence reads and WGS sequence reads were extracted from the samples. We compared statistical inferences performed using WGS data and functional pathways predicted from 16S rRNA gene sequences.

4.1.3 Results. We found that the distributions of predicted gene pathway assignments and the pathway assignments made using WGS sequence reads were broadly similar. We observed pronounced clustering by cage in the functional profiles of microbial communities both quantified with WGS sequences and predicted functional profiles.

However, there were numerous predicted gene pathways that were associated with cage effect in the predicted functional pathways which were not significant in the actual WGS reads.

4.1.4 Discussion. We conclude that while 16S rRNA sequence reads can be useful in viewing the overall distributions of functional content, these predicted functions tend to incur a high proportion of false positives when used to make statistical inferences.

4.2 Introduction

There has been considerable recent interest in determining the role that gut microbes play in driving disease and maintaining health in the host. Whole genome metagenome shotgun (WGS) sequencing – sequencing in which all DNA in a sample is used to generate sequences – is a commonly used method to study the functions of complex microbial communities. While informative, obtaining WGS sequences can be more costly than 16S rRNA sequencing[38], can require a large amount of input DNA and the analysis of WGS sequences can be time consuming. In this study, we measured the accuracy of functional profiles of gut microbiota predicted by 16S sequence reads and compared them to functional profiles generated using WGS sequence reads in a study investigating if there was a link between hepatic steatosis (HS) and dietary fructose[111].

Hepatic steatosis (HS) is a condition in which fat accumulates on the liver. HS has been associated with the development of diabetes. It has been hypothesized that one possible cause of HS is the leakage of bacteria from intestinal sites to surrounding organs, a process known as microbial translocation (MT)[112]. The increased use of dietary fructose, which has grown in recent decades via increased consumption of soft drinks has been associated with increased Type 2 diabetes[113]. The current study is part of an on-going series of studies to test the hypothesis that there is a possible link between HS and high dietary fructose consumption. We have previously published 16S data[111, 114] on a vervet (*Chlorocebus aethiops)* colony in which fructose was changed via a dietary manipulation and here present for the first time WGS sequencing from these samples.

Because the cost of WGS sequencing can be quite expensive in comparison to target sequencing a marker gene such as 16S rRNA, many studies only conduct 16S sequencing. However, 16S sequences only give insights about the phylogenetic diversity in a microbial community and not functional profiles of the microbial community. There have been efforts to predict the functional capabilities of a microbial community using 16S rRNA sequence reads. One example of a software program that makes these types of predictions is PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States)[39]. PICRUSt infers metagenomic functional content by taking operational taxonomical units (OTUs), which represent a collection of microbes that share some threshold of 16S sequence similarity, and mapping these OTUs representative sequences to previously sequenced bacterial genomes. The PICRUSt developers used data in part derived from the Human Microbiome Project (HMP)[6, 42] to test their algorithm. The creators of PICRUSt found that their algorithm had a high accuracy (Spearman R = 0.9*, p*-value < 0.001) when predicting the functions of the human microbiome community using data from the HMP. One goal of our analysis was to see how well this algorithm performs on a non-human primate system. We found that while the PICRUSt was reasonably accurate in predicting functional profiles, its predictions were less reliable when used for inference within linear models.

4.3 Methodology

4.3.1 Sample Description

10 vervet monkeys from the Wake Forest Primate Center that were involved in a previous study that determined dietary fructose caused liver injury in primates[111] were used for this study. For six weeks, half of the 10 monkeys were fed a diet that was high in fructose (24% of the caloric intake of the diet was from fructose) (HFr diet) and the other five were on a control diet (<0.5% of caloric intake of the diet was from fructose) (Chow diet). Stool samples were collected over the course of the study creating 52 matched WGS sequences and 16S rRNA samples. Six of the vervets were classified as "old" (more than 15 years old) while the remaining four were classified as "young" (less than 9 years old). For the duration of the study the monkeys were housed in 6 cages (two monkeys were in cages alone and the other four cages housed two monkeys each) (Table 4.1). We have previously used the 16S rRNA sequences that were generated from the 52 samples to determine associations between the makeup of the microbial communities of the monkeys and factors such as diet and age[114].

4.3.2 Classification of 16S sequences

In order to classify the bacteria in the gut of the vervets, we first clustered reads from 16S sequences into groups called operational taxonomic units (OTUs) using the program AbundantOTU[115]. To assign taxonomy, consensus sequences from each of the OTUs were ran against a 16S rRNA database called the Ribosomal Database Project (RDP)[104, 105] using the RDP classifier. To account for differences in raw counts across the samples, the taxonomic classification table was log normalized[17]. We then filtered out columns in which taxonomic classifications are only in 20% or less of the samples (Table

4.2). Multidimensional scaling (MDS) ordination was done with Bray-Curtis similarity using the vegan package in R

4.3.3 Assigning gene functions to whole-genome sequences

WGS sequence reads were generated from the 52 vervet stool samples. To ensure that only sequence reads originating from bacterial gene functions were used, vervet sequences were removed by running the reads against the vervet (*Chlorocebus aethiops*) genome using BLAT[116]. Next, the WGS reads were searched against the KEGG protein database[106, 107] using BLAST[108] to map the WGS reads to KEGG gene families. The BLAST hits were further investigated using HUMAnN (The HMP Unified Metabolic Analysis Network)[109]. The resultant pivot table was then log normalized using the normalization method described in Section 1.6.

To visualize similarities in the gene pathways in the community we conducted a MDS analysis on the KEGG gene pathway abundance tables using the "capscale" function in R's vegan package. We used the Bray-Curtis metric to define the dissimilarity matrix for the MDS.

We then predicted the microbial functions from the 16S reads using PICRUSt with the default parameters[25]. For each sample, we performed a nonparametric Spearman correlation on the list of KEGG gene pathway abundances from both WGS and PICRUSt to determine the accuracy of PICRUSt predicted functional abundances compared to WGS-derived functional abundances.

4.3.4 Predicting gene functions using PICRUSt

In order to predict the microbial functions from the 16S reads, we used the software package PICRUSt. PICRUSt works to predict metagenomic content of a microbial community using marker gene sequences. PICRUSt first uses a reference OTU tree and uses the tips of the tree whose gene content is known to help infer gene content of OTUs with unknown gene content using ancestral state reconstruction methods. From this step, PICRUSt creates a table with OTUs as rows and gene content predictions as columns called a "gene content predictions table". As an input, PICRUSt takes an OTU table derived from 16S rRNA sequences, multiplies the OTU table by the "gene content predictions table" to create a predicted gene family table using the OTU table.

In order to obtain KEGG pathways or KEGG modules, the predicted KEGG gene families from PICRUSt are then used as inputs for the HUMAnN program[109]. KEGG pathways are a quantification of molecular interactions and networks and KEGG modules are collection of functional units such as structural complexes in a living system[106, 107].

4.3.5 Statistical Analysis of gene pathways

4.3.5.1 MDS association models

We utilized marginal linear models to determine the associations between the MDS axes generated from KEGG gene pathway abundances using each of the following factors as fixed variables: age, diet type, days into the study in which the sample was taken. The "(1| Cage/Individual)" represents a term that tells the model that samples within the sample cage are correlated and within each cage samples from the same animal have an even higher correlation and the different cages will contribute a different value to the intercept (the "1" in "(1|Cage/Individual)" represents the intercept):

$$MDS\ axis\ or\ KEGG\ function = Days + Diet + Age + (1|Cage/Individual)\ + e, \text{\{Equation 4.1\}}$$

4.3.5.2 KEGG pathway association models

In order to test how the vervets' diet, age, cage assignment and time in the study all contribute to the variance of the KEGG pathways of the microbial communities of the vervets' guts we used marginal linear models (Equation 4.1). The terms for these models are same as the terms described above. These models were generated on both the PICRUSt predicted gene pathways and WGS-gene pathways.

4.3.5.3 PICRUSt statistical inference performance

To determine whether the cage effect in both PICRUSt and WGS derived gene pathways match we run a marginal model similar to that of Equation 4.1 but with the all the samples treated independently (Equation 4.2).

$$MDS\ axis\ or\ KEGG\ function = Days + Diet + Age + \ e, \text{ \{Equation 4.2\}}$$

We then run an ANOVA comparing Equations 4.1 and 4.2, which generates a p-value using the likelihood ratio test. The p-value that is generated for this ANOVA will be used to quantify the cage effect.

4.4 Results

4.4.1 Distribution of structure and functions of vervet gut microbiome similar to that of the human gut microbiome

We generated 16S rRNA and WGS sequence reads as a part of an on-going study of the effects of a diet high in fructose in a vervet colony[111]. As was the case for the Human Microbiome Project, samples showed a good deal of differences in their 16S profiles while WGS sequence reads mapped to the KEGG gene pathway database were much more consistent (Figure 4.1). We conclude that vervets, like humans, share a good deal of gene function even though this function is encoded in taxa that differ between individuals.

4.4.2 Clear clustering by cage and individual vervet but not by diet

Multidimensional scaling (MDS) of the counts of WGS reads assigned to KEGG gene pathways revealed, as previously reported[114] for the 16S sequences, there was a lack of clustering by diet (control or high fructose), age and date (Figure 4.2). We observed that there was a strong clustering by cage based on 16S taxonomic classifications (Figure 4.3) and this clustering by cage is also evidenced in clustering based on KEGG gene pathways (Figure 4.4). Statistical analysis found significant associations with cage for the first two MDS axes which collectively explains ~65% of the variation in the functional profiles of the samples (Table 4.5).

For each KEGG gene pathway that was identified in the WGS sequences reads, we ran marginal regression models to discover associations between the pathway abundances and diet, days, age and cage effects. We found that there were no KEGG gene pathways that were associated with amount of days that the vervets were on this study. We also found when we collapse KEGG gene pathways to a higher level that 77% (22/27) of these higher level gene pathways are significantly associated with cage at a 10% FDR (Table 4.6).

4.4.3 PICRUSt Predicted gene functions are consistent with WGS gene functions

In order to determine whether or not we can reproduce our WGS KEGG pathway findings using 16S sequences, we compared PICRUSt KEGG pathway predictions to WGS KEGG pathway profiles. In general, the KEGG gene pathways quantified using WGS sequences and PICRUSt predicted KEGG gene pathways were well correlated with Spearman coefficient averaging ~87% across the samples (Figure 4.5) (Table 4.7). The distribution of KEGG gene pathways of gene families predicted from PICRUSt was largely uniform across the samples much like that of the distribution of KEGG gene pathways calculated using the WGS sequences (Figure 4.5 A, B). As was the case for the WGS sequences, PICRUSt predicted KEGG gene pathways showed little significant differences due to the diet, time and age but a significant cage effect (Table 4.9).

4.4.4 PICRUSt cage effects in predicted gene functions are inconsistent with cage effects in WGS gene functions

For each of the WGS-derived and PICRUSt-predicted functions, an ANOVA test was performed to test for the strength of cage effects. For the KEGG gene pathways at the most specific level (level 3), 57.2% (83/145) of the WGS-derived functions had a significant association with cage at a 5% FDR threshold, while over 80% (136/168) of the PICRUSt predicted gene pathways showed significant differences. In comparing the *p*-values from an ANOVA test under the null hypothesis of no association with cage for the PICRUSt predictions and the WGS-derived functions, we found that there was a significant correlation between the *p*-values that were generated from PICRUSt predicted functions and *p*-values generated from WGS functions (Appendix B) (Figure 4.7). There were numerous gene pathways, however, that did not have an association with abundance and cage but were significantly associated with cage in PICRUSt predicted abundance. We noticed that while both “real” WGS sequences and PICRUSt predicted sequences can report cage effects, there was a tendency for PICRUSt to over predict differences in gene family abundances that are associated with cage.

4.5 Discussion

This study was part of an on-going series of studies designed to determine if there is a difference in the gut microbiota of vervets fed a diet high in fructose versus vervets fed a controlled diet. We found no significant difference in the microbial communities due to the increase of fructose in the diet of the vervets. Despite the lack of differences in taxa associated with the vervets' consumption of fructose this does not necessary mean that fructose has no impact on the gut microbiota of the vervets. The lack of signal in the microbiota associated with fructose could be explained by the small sample size of the vervets that were used for this study. Another explanation is that the gut microbiota of these animals could be quite resilient to short term changes in the diet, but a longer study might have found more differences. We also saw no significant differences in age of the vervets or the day in which the sample was taken. By contrast, there was a large association of the microbiota with cage, which is clearly a strong confounding variable in this study. We note that age is confounded with the family history of the monkeys as related animals tended to be co-housed; we did not attempt to separate these variables in our current analysis. In addition, for several years before the animals were placed in housing assignments for this study, they were historically housed in a total of only 3 cages. Again, we did not attempt to resolve the history of co-housing in our analysis.

A strong cage effect has also been observed in mouse models[11, 22]. Intriguingly, a potentially similar phenomenon has been observed in humans where patients in nursing homes have a distinct microbiome when compared to the general population [9], although this was not discussed as a "cage effect". We used our observed cage effect in the vervets to test the ability of PICRUSt, a program which infers functional content of microbial communities using 16S rRNA sequence read counts, to make inferences about differences in functional content due to different conditions (in our case the cage in which the vervet was housed). Testing this was possible because we had both the PICRUSt predicted functional content and functional content inferred from WGS sequence reads.

Through this test we found that PICRUSt functional predictions was highly correlated with the functional abundances that were inferred from WGS sequence reads. This suggests that PICRUSt can successfully be extended to non-human primate samples when determining the overall functional content of a microbiome. We also found that PICRUSt had some limitations. When used for hypothesis testing, PICRUSt generated gene pathways with *p*-values indicating a significant cage effect when these same pathways were not significant using abundances that were directly from WGS data. There were also cases when the PICRUSt predicted functions did not detect a cage effect when there was one according to KEGG pathway abundances measured using WGS sequence reads. This shows that although PICRUSt may be a powerful tool to use to find out the functions of a microbial community using 16S rRNA sequences, the predicted functional data from PICRUSt may not be very reliable for making statistical inferences. In fairness, the creators of PICRUSt warns users that if the Nearest Sequenced Taxon Index (NSTI) per sample is too high (≥ 0.15), then the results from PICRUSt could be of low quality. Our samples from vervets had an average NSTI of ~ 0.16. This may have impacted our results. This work suggests that due caution should be used in using tools like PICRUSt for inference in animal systems that are used as models for host/microbiome relationships such as mice, drosophila and non-human primates.

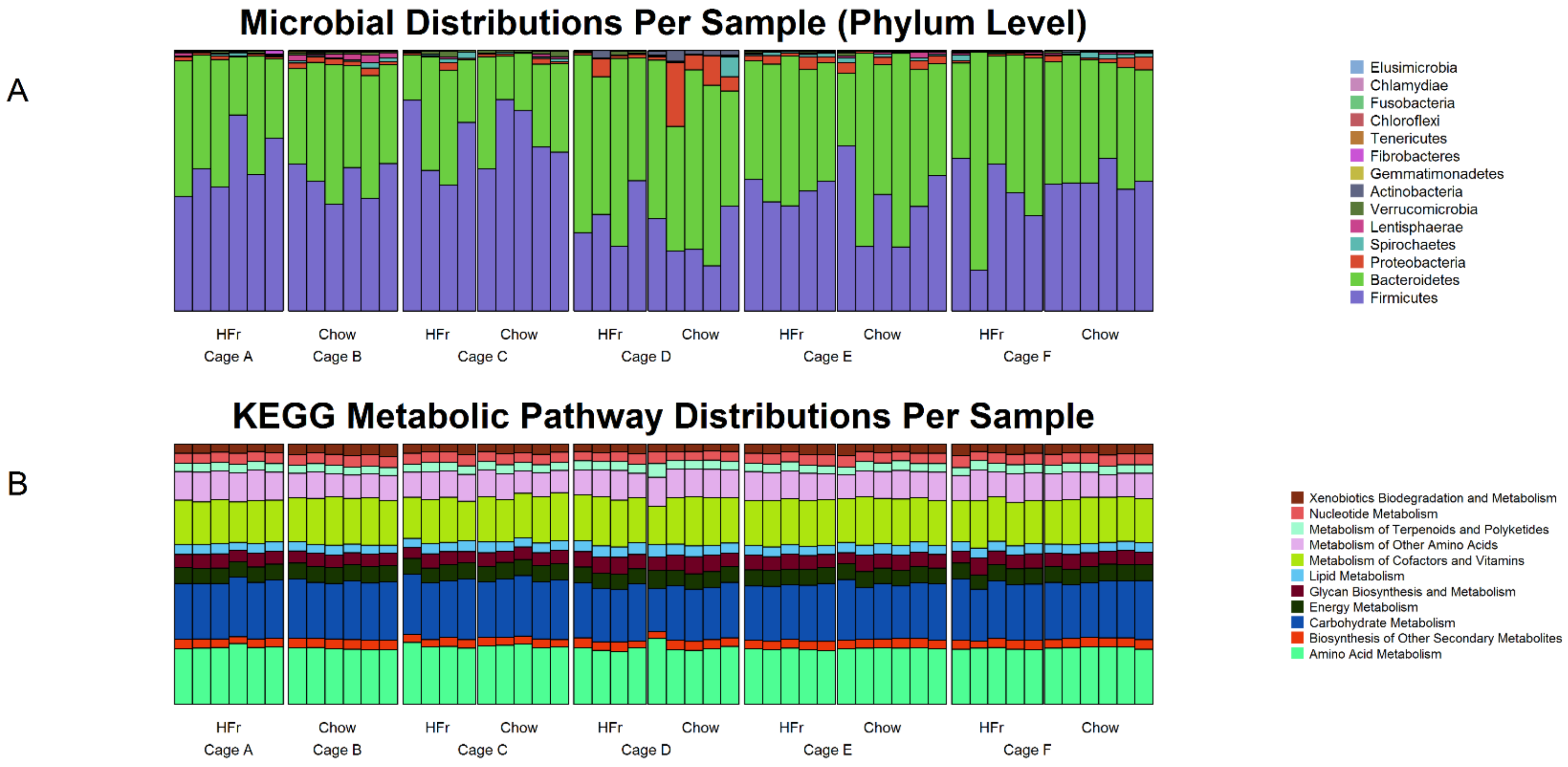

Figure 4.1: The profiles of microbial classifications differ among the individual vervets, however the distributions of KEGG metabolic pathways was similar across the samples.

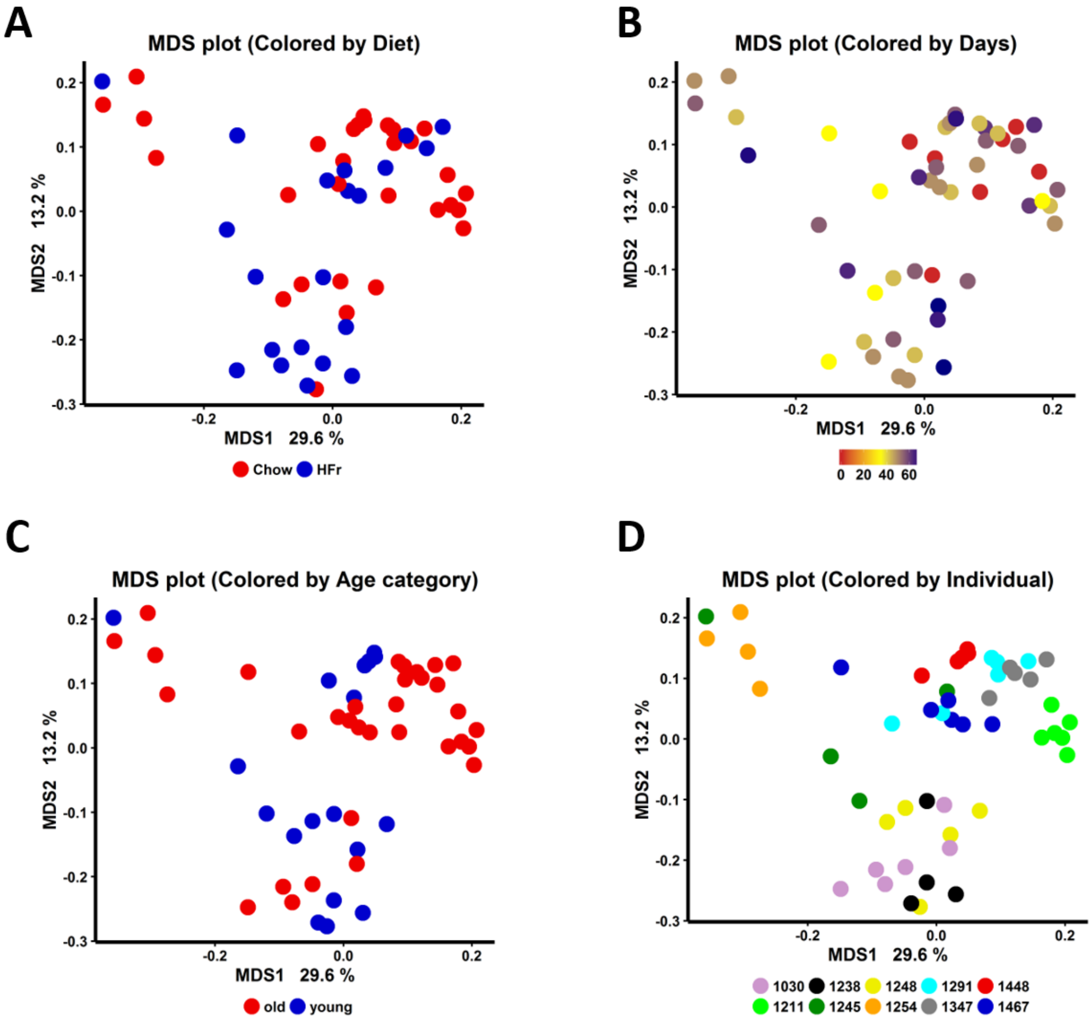

Figure 4.2: Multidimensional scaling (MDS) shows that while the age, time and diet has little differences, samples were different based on individuals. The x-label shows the amount of variation explained by the first MDS axes and the y-label is the amount of variation explained by the second MDS axes. The points represent each sample and are colored by diet (A), age (B), date (C), and individual (D).

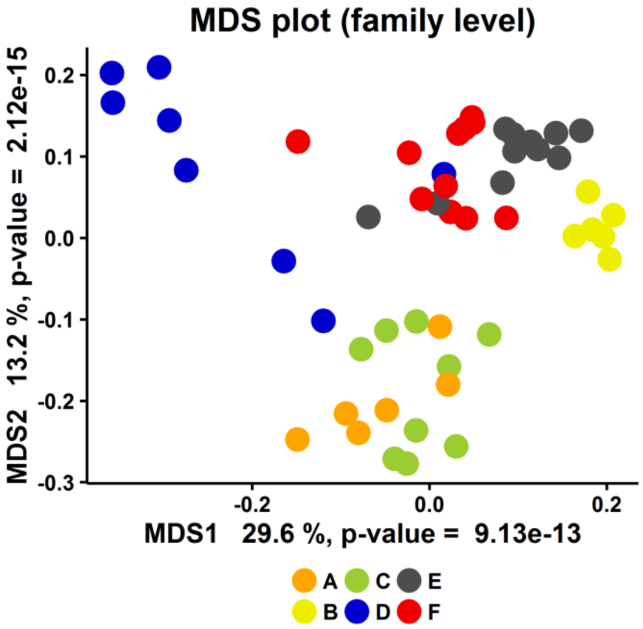

Figure 4.3: Multidimensional scaling (MDS) shows that samples were different based on cage.

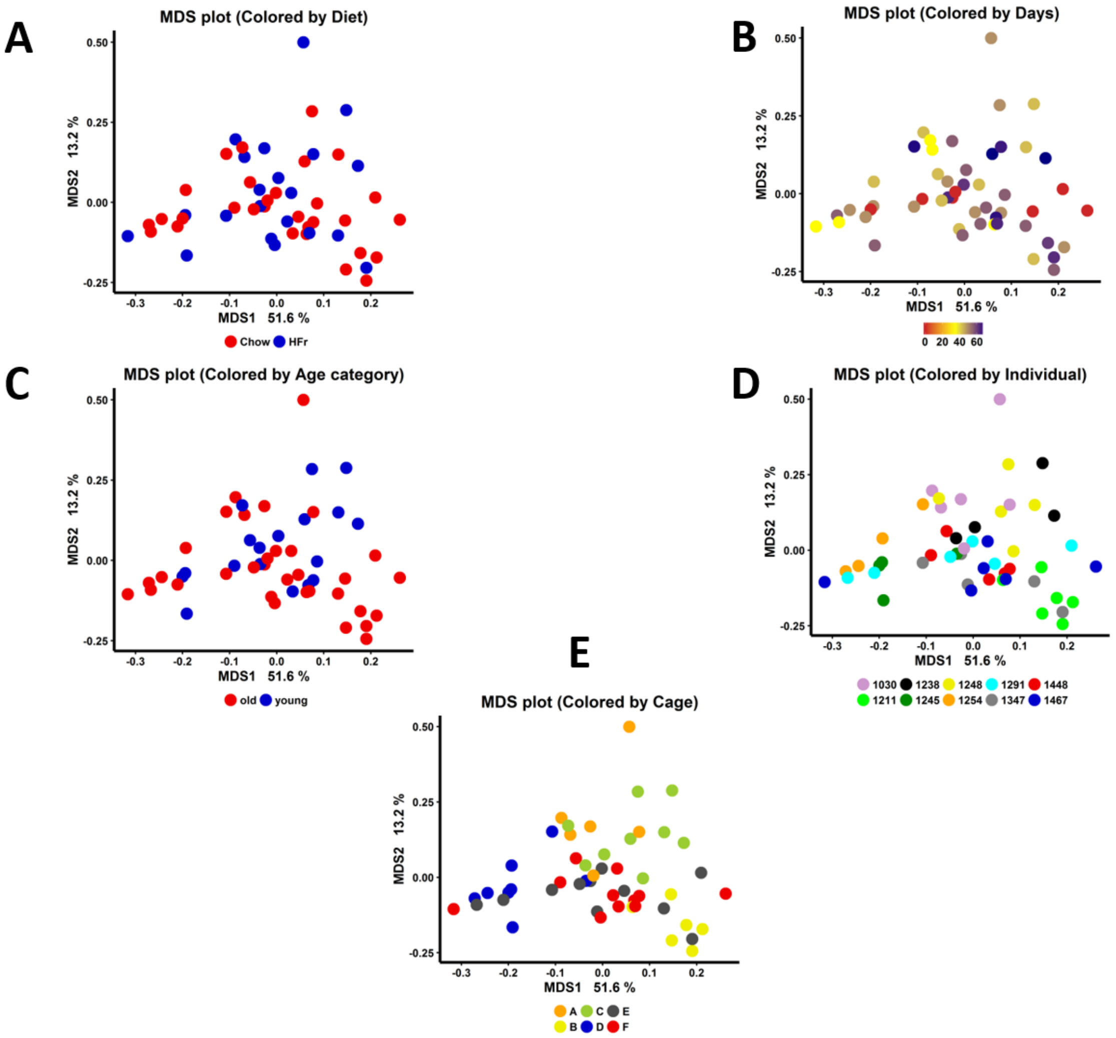

Figure 4.4: Multidimensional scaling (MDS) of WGS data shows that while the age, time and diet has little differences, samples were different based on individuals and cage. The x and y labels show the amount of variation explained by the first and second MDS axes, respectively. The points represent each sample and are colored by diet (A), age (B), date (C), individual (D) and cage (E).

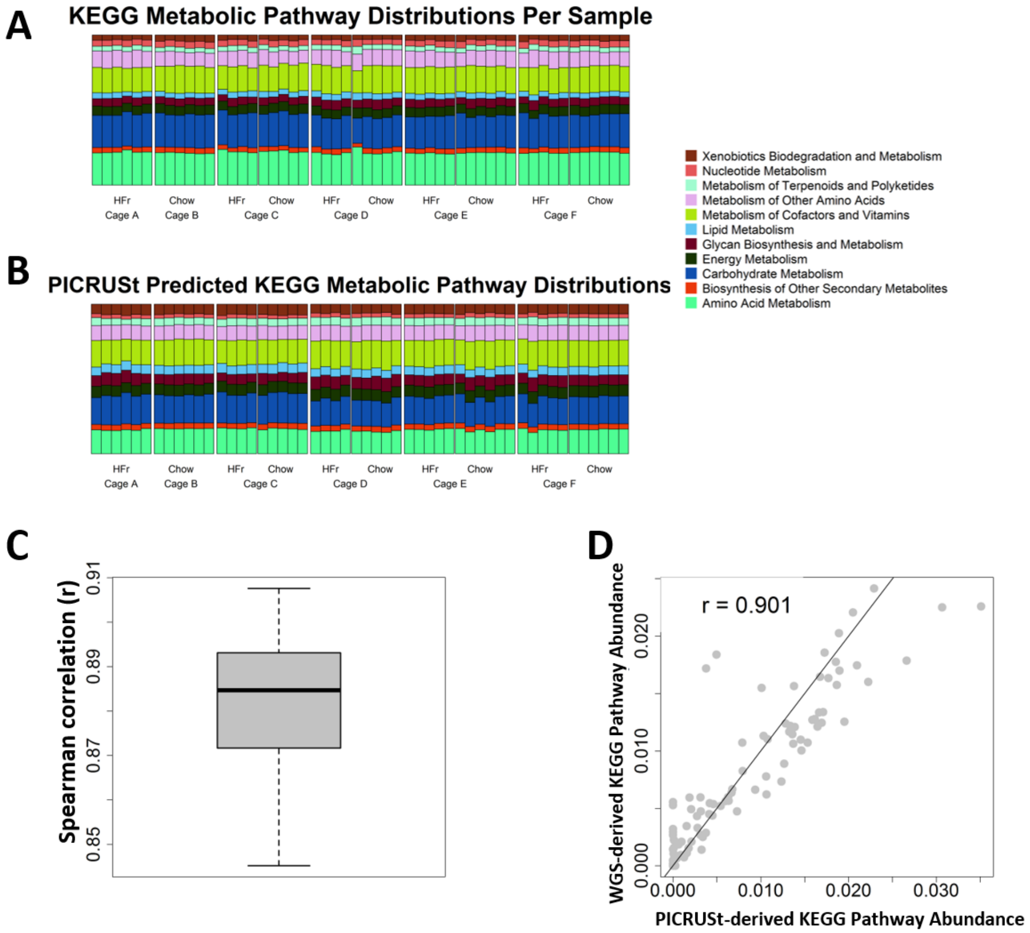

Figure 4.5: Correlation of KEGG metabolic pathway abundances predicted by PICRUSt and those calculated from WGS sequences reads were was fairly high. (A) Spearman correlation of predicted KEGG pathway abundances and KEGG pathway abundances calculated from WGS sequences across the 52 samples. (B) Scatterplot of PICRUSt-predicted KEGG metabolic pathway abundances and WGS-derived metabolic pathway abundances for one sample. The functional profiles (as shown by KEGG metabolic pathways) of the vervets (C) are very similar and likewise the functional profiles predicted by PICRUSt appears to be very similar (D).

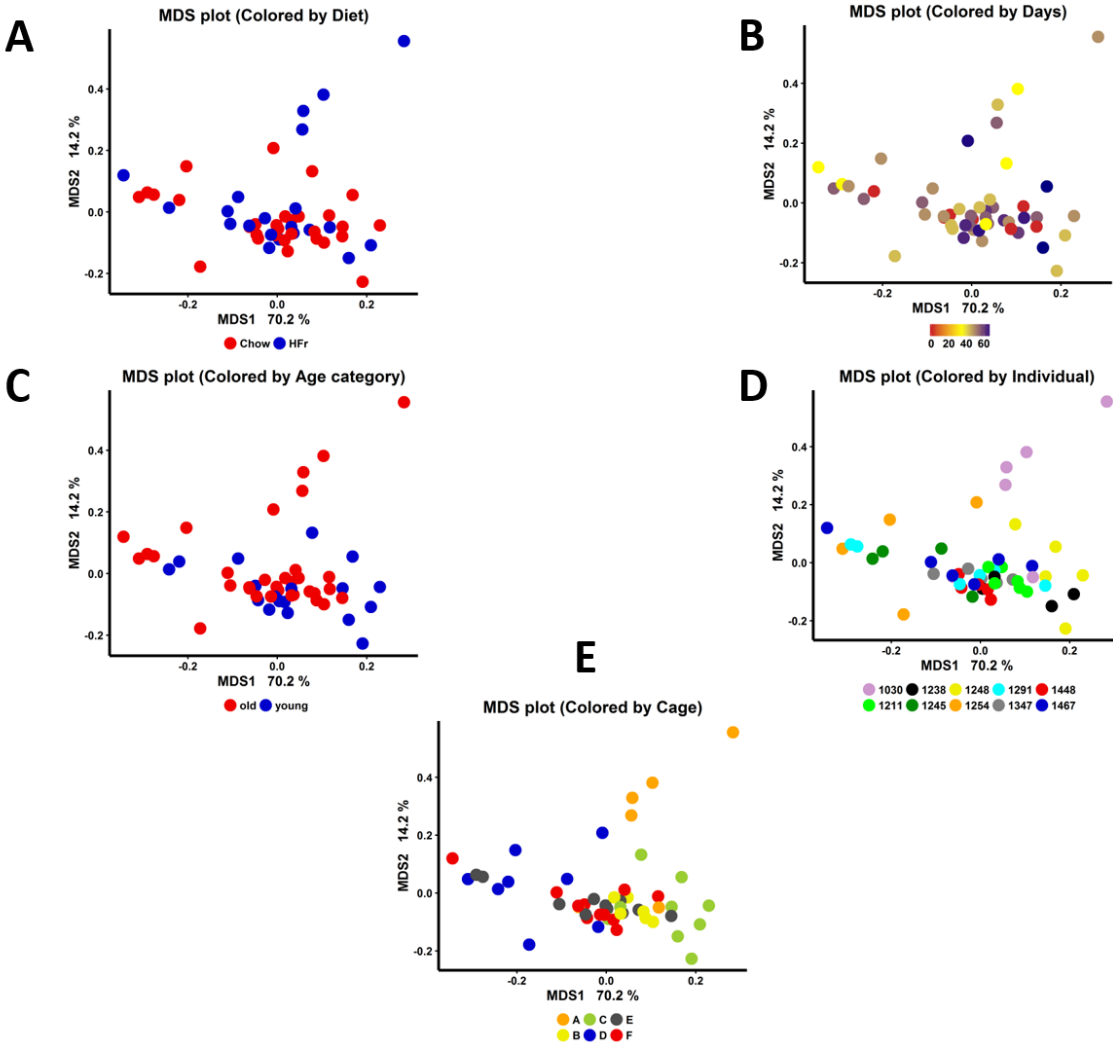

Figure 4.6: Multidimensional scaling (MDS) of PICRUSt data shows that while the age, time and diet has little differences, samples were different based on individuals and cage. The x and y labels show the amount of variation explained by the first and second MDS axes, respectively. The points represent each sample and are colored by diet (A), age (B), date (C), individual (D) and cage (E).

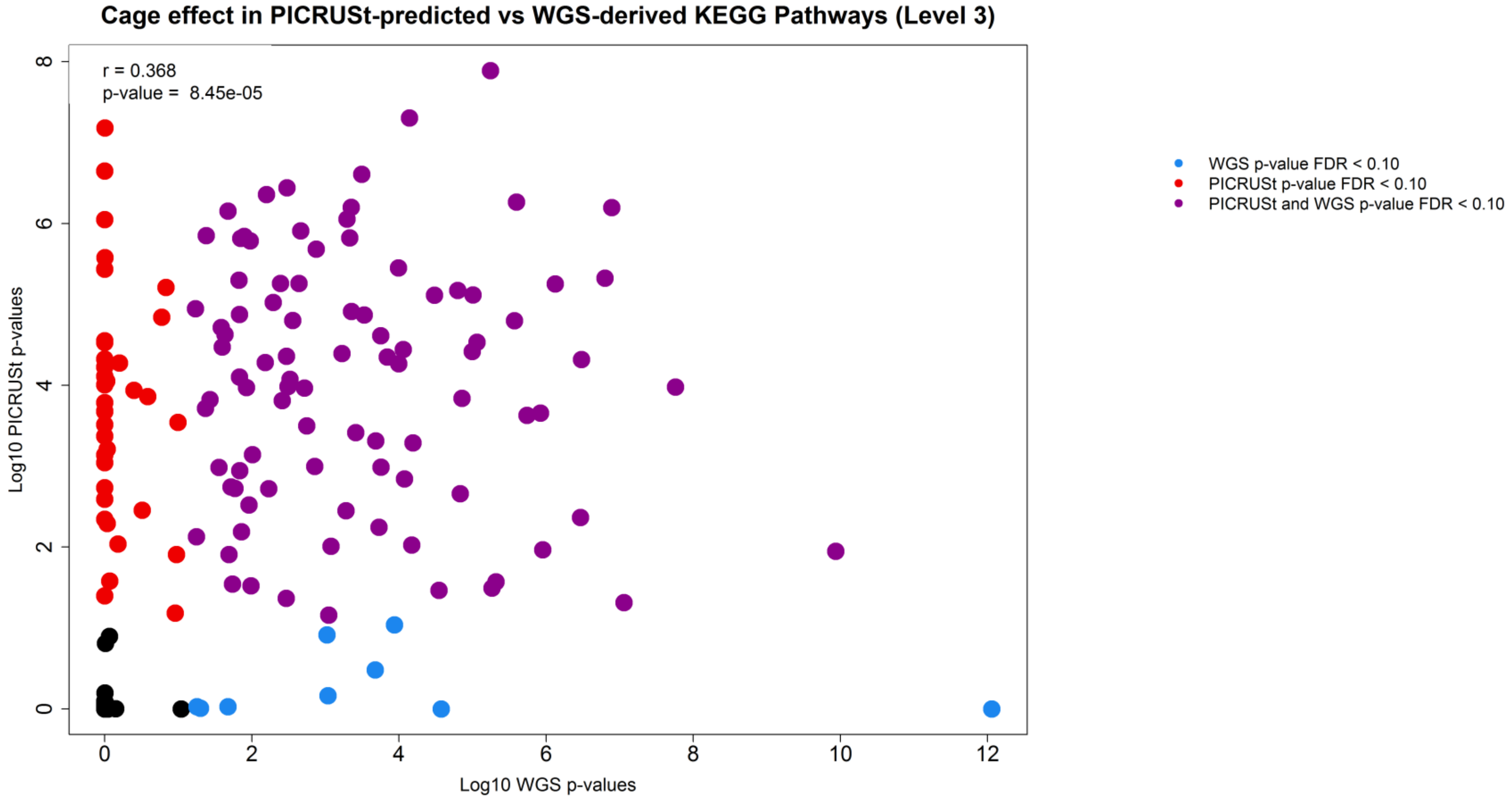

Figure 4.7: Results of using PICRUSt predicted pathway abundances for statistical inference significantly correlates with results of using WGS derived pathway abundances. An ANOVA test was conducted to determine if there was a difference in KEGG metabolic pathway abundances due to cage and also an ANOVA test was conducted for the PICRUSt predicted KEGG metabolic pathway abundances.

Table 4.1: Description of 52 paired 16S sequence samples and WGS sequence samples were generated.
🟥 = High- Fructose 🟩= Chow ■= No Sample Taken

Animal #	**Age Cat.**	**Cage**	**Baseline (Day -33)**	**Day 0**	**Day 9**	**Day 15**	**Day 22**	**At Biopsy (Day 27-35)**
1030	Old	A	🟩	🟥	🟥	🟥	🟥	🟥
1211	Old	B	🟩	🟩	🟩	🟩	🟩	🟩
1248	Young	C	■	🟩	🟩	🟩	🟩	🟩
1238	Young	C	■	■	🟥	🟥	🟥	🟥
1254	Old	D	🟩	■	🟩	🟩	🟩	🟩
1245	Young	D	🟩	■	■	🟥	🟥	🟥
1291	Old	E	🟩	🟩	🟩	🟩	🟩	🟩
1347	Old	E	🟩	■	🟥	🟥	🟥	🟥
1448	Young	F	🟩	■	🟩	🟩	🟩	🟩
1467	Old	F	🟩	🟥	🟥	🟥	🟥	🟥

Table 4.2: Statistics of 16S rRNA sequence reads after various filtering steps.

	Number of Samples	Number of OTUs	Total Number of Sequence Reads	Mean Reads per sample ± SD (SE)	Minimum reads per sample	Maximum reads per sample
16S reads generated	52		35,885,214	690,100.3 ± 104,617.42 (95,699.69)	417,228	874,594
After clustering into OTUs	52	4,562	32,334,816	621,823.4 ± 91,334.67 (86,231.39)	384,965	788,919
After filtering out OTUs in less than 20% of	52	1,849	32,222,426	619,662.0 ± 91,208.30 (85,9331.66)	376,904	787,478

Table 4.3: 16S Results of an ANOVA that tests the null hypothesis that there is no association between cage and coordinate for the first 10 coordinates at the family taxonomic level.

MDS Axis	**% of variation explained**	**Diet *p*-value**	**Age *p*-value**	**Time *p*-value**	**Cage *p*-value**
1	29.638	0.068	0.9259	0.44	4.07E-12
2	13.216	0.3453	0.7415	0.893	4.07E-12
3	8.85	0.9683	0.9259	0.334	1.00E-05
4	7.705	0.1574	0.0649	0.893	7.26E-06
5	5.737	0.078	0.9259	0.878	0.000173
6	4.576	0.068	0.9259	7.95E-07	0.01665
7	4.108	0.4828	0.7222	0.415	1.00E-05
8	3.44	0.7213	0.7415	0.483	0.085795
9	2.902	0.0682	0.9914	0.893	0.000269
10	2.533	0.9269	0.7661	0.878	0.154449

Table 4.4: 16S Results of a marginal linear model for all taxa at the phylum level. *p*-values are adjusted at a 10% false discovery rate.

Phylum classification	**Diet *p*-value**	**Age *p*-value**	**Time *p*-value**	**Cage Effect *p*-value**
Actinobacteria	0.5115	0.181	0.846	2.22E-06
Bacteroidetes	0.8909	0.362	0.846	0.001422
Chlamydiae	0.8909	1	0.846	0.61765
Chloroflexi	0.9491	1	0.846	0.656927
Elusimicrobia	0.9491	1	0.898	0.977519
Fibrobacteres	0.5115	0.635	0.846	0.42251
Firmicutes	0.9367	0.143	0.846	4.61E-05
Fusobacteria	0.9491	1	0.846	0.656927
Gemmatimonadetes	0.8909	1	0.846	0.0019
Lentisphaerae	0.0936	0.515	0.112	0.001422
Proteobacteria	0.0936	0.181	0.112	0.005008
Spirochaetes	0.3152	1	0.596	0.019451
Tenericutes	0.8909	1	0.846	0.656927
Verrucomicrobia	0.8909	1	0.846	0.000921

Table 4.5: WGS sequences classified to KEGG pathways and ordinated. *P*-values from marginal model of the first 10 MDS axes. *P*-values are corrected at a 10% false discovery rate.

MDS Axis	**% of variation explained**	**Diet *p*-value**	**Age *p*-value**	**Time *p*-value**	**Cage *p*-value**
1	51.583	0.8	0.9631	0.712	0.0107
2	13.194	0.719	0.8652	0.641	2.72E-05
3	6.613	0.154	0.1162	0.712	0.6864
4	4.974	0.429	0.8652	0.972	0.4284
5	3.267	0.154	0.5155	0.896	0.4284
6	2.938	0.8	0.8652	0.972	0.4284
7	2.317	0.154	0.0995	0.397	0.6864
8	2.145	0.278	0.9332	0.986	0.7623
9	1.916	0.8	0.8652	0.896	0.8162
10	1.51	0.8	0.8652	0.712	0.8104

Table 4.6: WGS KEGG pathway (level 2) Results of a linear marginal model for all taxa at the phylum level. *P*-values are adjusted at a 10% false discovery rate.

KEGG pathway classification	**Diet *p*-value**	**Age *p*-value**	**Time *p*-value**	**Cage *p*-value**
Metabolism of Terpenoids and Polyketides	0.92	0.9736	0.405	0.00882
Metabolism of Cofactors and Vitamins	0.549	0.9736	0.405	0.05732
Transcription	0.991	0.9736	0.88	0.05386
Nucleotide Metabolism	0.92	0.9736	0.88	0.0033
Immune System	0.991	0.9736	0.88	0.49502
Xenobiotics Biodegradation and Metabolism	0.991	0.8992	0.88	0.00102
Membrane Transport	0.991	0.9736	0.88	0.00658
Cell Growth and Death	0.991	0.9736	0.405	0.02692
Folding Sorting and Degradation	0.92	0.9736	0.88	0.03211
Transport and Catabolism	0.988	0.9736	0.971	0.12146
Excretory System	0.991	0.9736	0.981	0.99933
Neurodegenerative Diseases	0.92	0.0304	0.88	0.00934
Metabolism of Other Amino Acids	0.075	0.0304	0.88	0.00882
Glycan Biosynthesis and Metabolism	0.991	0.9736	0.88	0.00102
Lipid Metabolism	0.905	0.8992	0.145	6.08E-06
Cell Motility	0.991	0.9736	0.88	0.01134
Translation	0.991	0.9736	0.981	0.00343
Infectious Diseases	0.991	0.1405	0.405	0.00882
Signal Transduction	0.905	0.9736	0.88	0.00263
Replication and Repair	0.402	0.9736	0.88	0.0421
Carbohydrate Metabolism	0.991	0.3926	0.405	0.03488
Environmental Adaptation	0.991	0.9433	0.88	0.11217
Amino Acid Metabolism	0.991	0.8992	0.706	0.0421
Biosynthesis of Other Secondary Metabolites	0.301	0.8992	0.88	0.00698
Endocrine System	0.991	0.9433	0.405	6.32E-06
Digestive System	0.999	0.9736	0.981	0.36216
Energy Metabolism	0.075	0.9736	0.971	0.03211

Table 4.7: Correlation of KEGG metabolic pathway abundances predicted by PICRUSt and those calculated from WGS sequences reads was higher as higher levels. Spearman correlation of predicted KEGG pathway abundances and KEGG pathway abundances calculated from WGS sequences across the 52 samples.

	Average Accuracy (%)	**Maximum Accuracy (%)**	**Minimum Accuracy (%)**
KEGG Gene Pathways (Level 1)	100.00	100.00	100.00
KEGG Gene Pathways (Level 2)	94.31	96.52	91.58
KEGG Gene Pathways (Level 3)	88.94	91.18	86.50
KEGG Gene Pathways (Level 4)	86.85	89.76	83.94

Table 4.8: PICRUSt-predicted KEGG pathways (level 2) were ordinated and here we show *p*-values from marginal model of the first 10 MDS axes. *P*-values are corrected at a 10% false discovery rate.

MDS Axis	**% of variation explained**	**Diet *p*-value**	**Age *p*-value**	**Time *p*-value**	**Cage *p*-value**
1	70.189	0.977	0.85	0.847	0.0156
2	14.241	0.611	0.675	0.847	0.0361
3	5.665	0.611	0.224	0.847	0.1915
4	2.501	0.977	0.667	0.847	0.0201
5	2.152	0.977	0.85	0.847	0.0608
6	0.857	0.779	0.85	0.847	0.6344
7	0.785	0.863	0.85	0.847	0.0763
8	0.576	0.977	0.85	0.847	0.7123
9	0.531	0.611	0.943	0.847	0.0505
10	0.356	0.977	0.85	0.847	0.825

Table 4.9: PICRUSt-predicted KEGG pathway (level 2) *p*-values from a linear marginal model. *P*-values are adjusted at a 10% false discovery rate.

KEGG pathway level 2 classification	**Diet *p*-value**	**Age *p*-value**	**Time *p*-value**	**Cage Effect *p*-value**
Metabolism of Terpenoids and Polyketides	0.99	0.98	0.17762	0.000173
Metabolism of Cofactors and Vitamins	0.99	0.531	0.14314	0.000173
Transcription	0.99	0.531	0.14314	0.000724
Nucleotide Metabolism	0.99	0.591	0.03927	6.67E-05
Immune System	0.99	0.975	0.12776	0.000368
Xenobiotics Biodegradation and Metabolism	0.99	0.531	0.17762	0.000328
Membrane Transport	0.99	0.337	0.26813	0.164971
Cell Growth and Death	0.99	0.975	0.23321	0.000398
Folding Sorting and Degradation	0.99	0.531	0.14314	0.064925
Transport and Catabolism	0.99	0.017	0.58872	0.854975
Excretory System	0.99	0.226	0.13312	0.50562
Neurodegenerative Diseases	0.99	0.577	0.45842	0.000398
Metabolism of Other Amino Acids	0.99	0.534	0.0658	0.000398
Glycan Biosynthesis and Metabolism	0.99	0.226	0.12776	0.079297
Lipid Metabolism	0.99	0.324	0.1001	0.003907
Cell Motility	0.99	0.975	0.17223	0.000206
Infectious Diseases	0.99	0.591	0.00197	2.11E-06
Translation	0.99	0.975	0.17223	0.000368
Immune System Diseases	0.99	0.26	0.14314	0.164971
Signal Transduction	0.99	0.975	0.14314	0.000398
Replication and Repair	0.99	0.98	0.17223	0.000208
Carbohydrate Metabolism	0.99	0.531	0.14314	0.000336
Environmental Adaptation	0.99	0.577	0.21116	0.002884
Amino Acid Metabolism	0.99	0.531	0.0093	5.00E-05
Biosynthesis of Other Secondary Metabolites	0.99	0.531	0.57969	0.245539
Endocrine System	0.99	0.151	0.12776	0.003965
Digestive System	0.99	0.577	0.06809	0.000724
Energy Metabolism	0.99	0.975	0.17762	0.000234

CHAPTER 5: IMPACT OF FRUCTOSE AND GLUCOSE ON THE GUT MICROBIOME OF RATS

5.1 Abstract

5.1.1 Background. While the obesity rise in the United States is well correlated with the rise of the used of high-fructose corn syrup in the Western diet, it is unknown whether this relationship is causal. In this chapter, we sought to determine the impact of varying sugar solutions of glucose and fructose on the gut microbiome of rats.

5.1.2 Methods. We fed rats sugar solutions that were either equal parts glucose and fructose, mostly glucose or mostly fructose. We generated 16S rRNA sequences from extracted stool samples to measure the composition of the gut microbiota of these rats as well as a no-sugar control group.

5.1.3 Results. We observe that the microbiota of rats that consumed sugar had a distinct microbiome from those given water instead of a sugar solution. Despite differences in the microbiota associated with the consumption of sugar, the type of sugar that was consumed appeared to have little association with the composition of the gut microbiota.

5.1.4 Discussion. Our results suggest that sugar can profoundly impact microbial community composition in juvenile rats, but the type of sugar (glucose/fructose ratio) does not appear to have a significant impact.

5.2 Introduction

The increase of dietary fructose in the American diet has coincided with the country's growing obesity epidemic. The United States has a higher per-capita intake of fructose than any other country[30]. A large portion of the fructose that is introduced in our diet is in the form of soft drinks sweetened with high-fructose corn syrup (HFCS), a liquid sweetener comprised of fructose and glucose and contains varying concentrations of fructose/glucose combinations. This increase in fructose associated with HFCS can lead to metabolic diseases and other problems such as fatty liver disease, obesity, and diabetes[29].

In collaboration with Dr. Michael Goran of the University of Southern California School of Medicine, 42 rats, which were housed individually, were used for this study. These rats were stratified into four groups that were given differing amounts of sugar. We found, independent of weight gain, that rats fed any of the sugar solutions had distinct microbial communities from the control group, but that microbial community composition was insensitive to the type of sugar fed to the rats.

5.3 Methodology

5.3.1 Experimental design

Forty-two juvenile, male Sprague Dawley rats (Envigo; PND 26; 50-70g) were housed individually in standard conditions with a 12:12 light/dark cycle and were classified into four groups based on solution feeding of: 1) 35% fructose and 65% glucose, 2) 65% fructose and 35% glucose, 3) 50% fructose and 50% glucose and 4) control (no sugar). For each of the sugar groups, the concentration of total sugar in solution was 11% w/v (comparable to sugar-sweetened beverages (SSB)s typically consumed by humans) in reverse osmosis-filtered water. In addition to sugar solutions (or an extra water bottle for the control group), rats were given access to standard chow and water *ad libitum*. Food intake, solution intake and body weights were monitored thrice weekly. After 6 weeks in respective conditions, feces were collected from the animals according to the following methods: each animal was placed in a sterile cage and gently restrained while lifting the tail until defecation occurred. Feces was immediately placed into dry ice and stored at -80 °C until time of processing for RNA sequencing. All experiments were performed in accordance with the approval of the Animal Care and Use Committee at the University of Southern California.

A separate group of male Sprague Dawley rats (n=42, PND 26; 50-70g) were housed individually in standard conditions with a 12:12 light/dark cycle and were classified into four groups in an identical design to cohort 1. After 6 weeks in respective conditions, body weights, chow intake, and sugar intake were similar to cohort 1 (data not shown). Body composition was assessed using a Bruker NMR minispec LF 90II (Bruker Daltonics Inc., Billerica, MA, USA). Adiposity index was calculated as fat mass (g)/lean mass (g) *100.

5.3.2 Taxonomic Classification of 16S rRNA gene sequence reads

Fecal microbiome populations were identified using next-generation high-throughput sequencing of the V3-V4 variable region of 16S rRNA (Vaiomer SAS, Labege, France). Genomic DNA was isolated and collected from fecal samples and DNA concentrations were determined using UV spectroscopy (Nanodrop2000, ThermoScientific). PCR amplification was done using 16S universal primers targeting the V3-V4 region of the bacterial 16S ribosomal gene (Vaiomer universal 16S primers), with a joint pair length encompassing 476 base pair amplicon tanks to 2 x 300 paired-end MiSeq kit V3. The detection of sequencing fragments was performed using MiSeq Illumina® technology. FastQ files were generated at the end of the run to perform the quality control and filtering. The 16S targeted sequences were then clustered of into OTUs before taxonomic assignment and analyzed using the bioinformatics pipeline as described [117, 118]. The paired sequence reads were then assigned taxonomy using Ribosomal Database Project (RDP) classifier [104] using a boostrap cutoff of 80%. Reads were also clustered at 97% similarity to obtain operational taxonomic units (OTUs) using the script "pick_de_novo_otus.py" from the QIIME software with default parameters [103]. The OTUs and taxonomy classifications were tabulated and to account for differences in raw counts across the samples, the tables were log normalized. Multidimensional scaling (MDS) was performed on the tables using the "capscale" function of the R statistical software package "vegan" [119] with Bray-Curtis dissimilarity.

5.3.3 Differences in microbiota of samples given the type of sugar consumed

In order to determine differences in classifications between rats fed sugar versus water (control), we used the following model:

$$\text{Abundance of bacteria}_{\text{i}} = \text{SugarVsControl} + \text{e}, \text{\{Equation 5.1\}}$$

The term "Abundance of bacteriai" represents the log-transformed normalized counts assigned to bacteria at a given taxonomic level. We determine the significance of differences in abundance of bacteria classified at a given taxonomic level relating to the type of sugar consumed using the following model:

$$\text{Abundance of bacteria}_{\text{i}} = \text{FructoseFraction} + \text{e}, \text{\{Equation 5.2\}}$$

where the fructose fraction was a quantitative variable ranging from 35 to 65. Statistical models were only built for "non-rare" taxa, which were present in at least 25% of all samples.

5.3.4 Relationship between bacteria and biomarkers

In addition to determining whether there are differences in the microbiome with respect to overall sugar intake or monosaccharide ratio, we examined whether bacterial type could explain differences in body weight (grams) or energy intake (kCal). A series of linear models were built to examine the association with these variables, which we will call intake variables:

$$\text{Abundance of bacteria}_{\text{i}} = \text{SugarVsControl } + \text{IntakeVariable} + \text{IntakeVariable} * \text{SugarVsControl} + \text{e}, \{\text{Equation 5.3}\}$$

One model was built for each combination of the Abundance of Bacteria (the log-normalized counts at a phylogenetic level) and the intake variables (body weight (grams), energy intake (kCal)).

5.3.5 Differential analysis of bacterial taxa using LEfSe

To identify bacterial taxa that were differentially abundant in the different dietary groups, Linear Discriminant Analysis Effect Size (LEfSe) analysis was performed on the RDP classification tables using the online Galaxy interface. Using the LEfSe algorithm, bacterial taxa that were differentially abundant in pairwise analysis of dietary groups were first identified and tested using the Kruskal-Wallis test. The identified features were then subjected to the Linear Discriminant Analysis (LDA) model with a threshold logarithmic LDA score set at 3.0 and ranked. Respective cladograms were generated with genus at the lowest level.

5.4 Results

In order to determine the effects of sugar on the microbial community, we performed an experiment with 42 individually housed rats in which 32 were given sugar solutions with varying glucose/fructose combinations and 10 received no sugar solutions and were used as controls. Fecal samples were taken from the animals and 16S rRNA sequencing on the Illumina platform was performed to characterize the microbial community. The RDP classifier was used to assign taxonomy to the 16S rRNA sequence reads and QIIME (Table 5.1) was used to cluster the sequence reads into operational taxonomic units (OTUs).

5.4.1 Effects of dietary sugar on abundance of fecal microbiota at different taxonomic levels

Results from our multidimensional scaling analysis (MDS) on the taxonomic classification tables at all phylogenetic levels represent a summary of gut microbial composition (Figure 5.1). Rats fed the sugar solution (red) compared to water (blue) had distinct clustering patterns (Figure 5.1 A-F). In order to determine whether similar clustering occurred as a function of monosaccharide ratio in the sugar solutions, we performed MDS analyses on rats that were fed either 35% fructose: 65% glucose, 50%F:50%G or 65%F:5%G. There was no clear separation based on the monosaccharide ratio of the sugar solutions given (Figure 5.1 G-L). The distribution of *p*-values derived from t-tests performed separately for each family show that about one-quarter of non-rare bacteria at the family level were significantly different between samples from rats given a sugar solution and control samples at a 10% false discovery rate (FDR) (Table 5.2) (Figure 5.2). By contrast, *p*-values derived from linear models with glucose/fructose as an independent variable were approximately uniform (Figure 5.3); None of the family-level bacteria had a difference in abundance with respect to sugar group at a FDR threshold of 10%. We conclude that the presence or absence of sugar has a large impact on the microbial community, but the type of sugar (glucose vs. fructose) did not produce significant differences in our experiments.

In order to further visualize our results within a phylogenetic context, pairwise comparisons were made comparing sugar with control at each phylogenetic level (Figure 5.4) with the program LeFSe (see methods). At the phylum level, *Proteobacteria* and *Actinobacteria* were elevated in all sugar groups compared with controls. At the class level *Actinobacteria* and *Bacilli* (of the phylum *Actinobacteria* and *Firmicutes*, respectively) were significantly elevated by sugar, as were *Alpha-, Beta-, and Gamma- Proteobacteria* (of the phylum *Proteobacteria*). Bacteria of the order *Lactobacillales, Actinobacteridae, Burkholderiales,* and *Enterobacteriales* were significantly elevated by dietary sugar. Under our LeFSe analysis, many taxa were significantly different between sugar and control at the family level, for example *Clostridiaceae_1*, *Lactobacillaceae, Rikenellaceae, Porphyromonadaceae, Bacteroidaceae, Bifidobacteriales, Sutterellaceae,* and *Enterobacteriaceae* were elevated by sugar, whereas *Prevotellaceae, Ruminococcaceae*, and *Lachnospiraceae* were reduced due to sugar consumption. At the genus level, *Prevotella* and *Lachnospiracea (incertae sedis)* were reduced by sugar consumption, whereas *Bacteroides, Alistipes, Lactobacillus, Clostridium (sensu stricto), Bifidobacteriaceae* and *Parasutterella* were all significantly elevated by sugar consumption (*p*-value < 0.05 and false discovery rate < 10% after correcting for multiple comparisons).

5.4.2 Relationship of fecal microbiota to body weight and energy intake

To determine how members of the microbial community are associated with body weight and calorie intake, we executed a series of linear regression models comparing these intake variables to log-normalized adjusted counts. Using Equation 5.3 as described in the methods, at a 10% FDR threshold, there were no significant associations with body weight or food intake to any member of the microbial community and there was no association with any of the interaction terms at any phylogenetic level. Likewise, the distribution of *p*-values for body weight or calorie intake generated by Equation 5.3 produced near-uniform *p*-values suggesting little association (Figure 5.5).

5.5 Discussion

This study aimed to determine the impact that varying fructose content in sugar solutions on the gut microbiome of rats. We found that while there was no association between the type of sugar consumed and the composition of the gut microbiota, there are some bacteria that distinguish between rats that consumed sugar and those who did not. From this we can conclude that despite the type of sugar that these rats consumed did not matter, but the presence of absence of sugar made a large difference in microbial community composition. This is significant to because as of lately there have been more soft drinks that are advertising that they are replacing their high-fructose corn syrup with "pure cane sugar" and while this may have an effect on other bodily functions, we find that this will not likely have an effect on the microbiome.

We did not find many interactions between measurements such as food intake (data not shown), and body weight with sugar versus control diet status that were associated with microbes. The microbes that we find that are higher in rats who consumed sugar are commonly associated with dysbiosis including families belonging to Proteobacteria including Burkholderiales and Enterobacteriales. Further studies could focus on the use of WGS sequencing to determine the KEGG gene pathways that could be associated with sugar intake. While we believe there will be little to no difference in the overall function of the gut microbiome of rats by the type of sugar solution, it would be useful to determine if there are any KEGG gene pathways that are significantly different in rats fed glucose vs. fructose. As part of my post-doctoral work, I will be involved with a study in which we will determine if there are differences in the gut microbial communities of obese human adolescents who are on a normal diet versus those who are on a reduced sugar diet. This would give us a better understanding of the role of gut microbiota and how it relates to sugar intake in human populations.

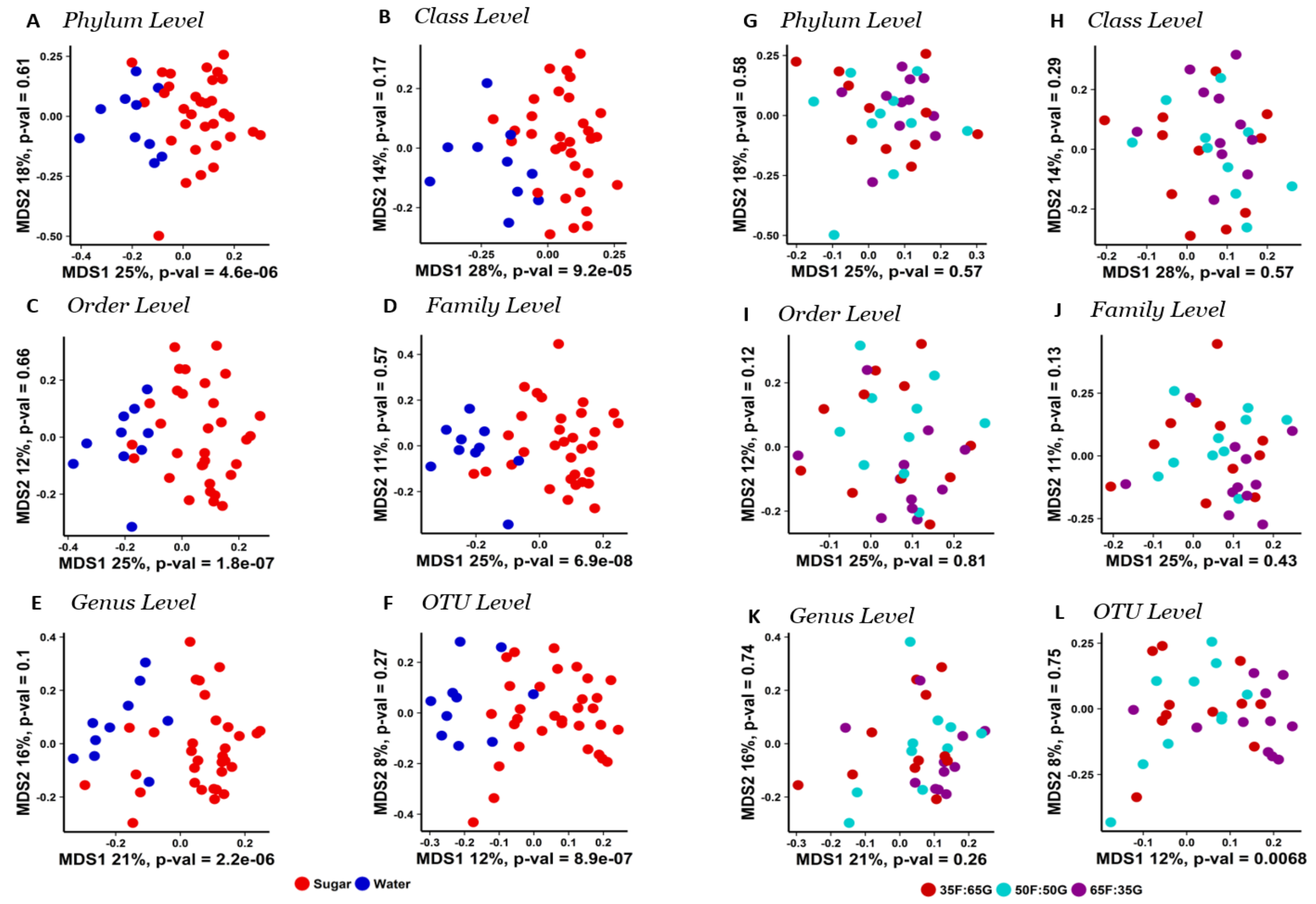

Figure 5.1: Multidimensional scaling (MDS) reveals that while there is a strong separation between samples given sugar solution and control samples. This pattern exists at all phylogenetic levels from phylum to OTU.

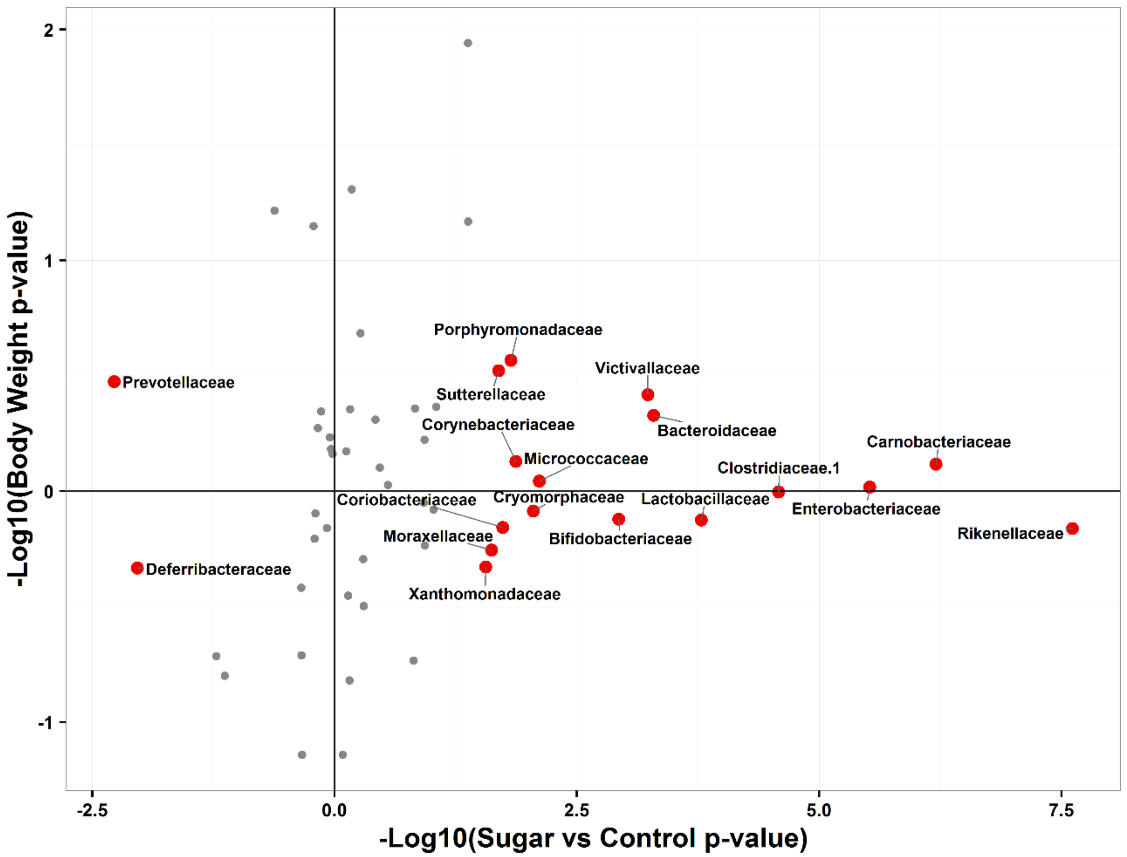

Figure 5.2: Numerous bacteria classified at the family are shows significant separation between sugar and control but only one shows significant separation by the type of sugar solution. The points in red denotes a Benjamini-Hochberg corrected *p*-value < 0.10 for separation by sugar versus control samples and no points were significantly separated by type of sugar solution consumed. *P*-values for the x-axis were generated by Equation 5.2 and are negative if the mean bacterial abundance in control samples is higher than the mean bacterial abundance in sugar samples and positive if the reverse is true. *P*-values for the y-axis were generated by Equation 5.2

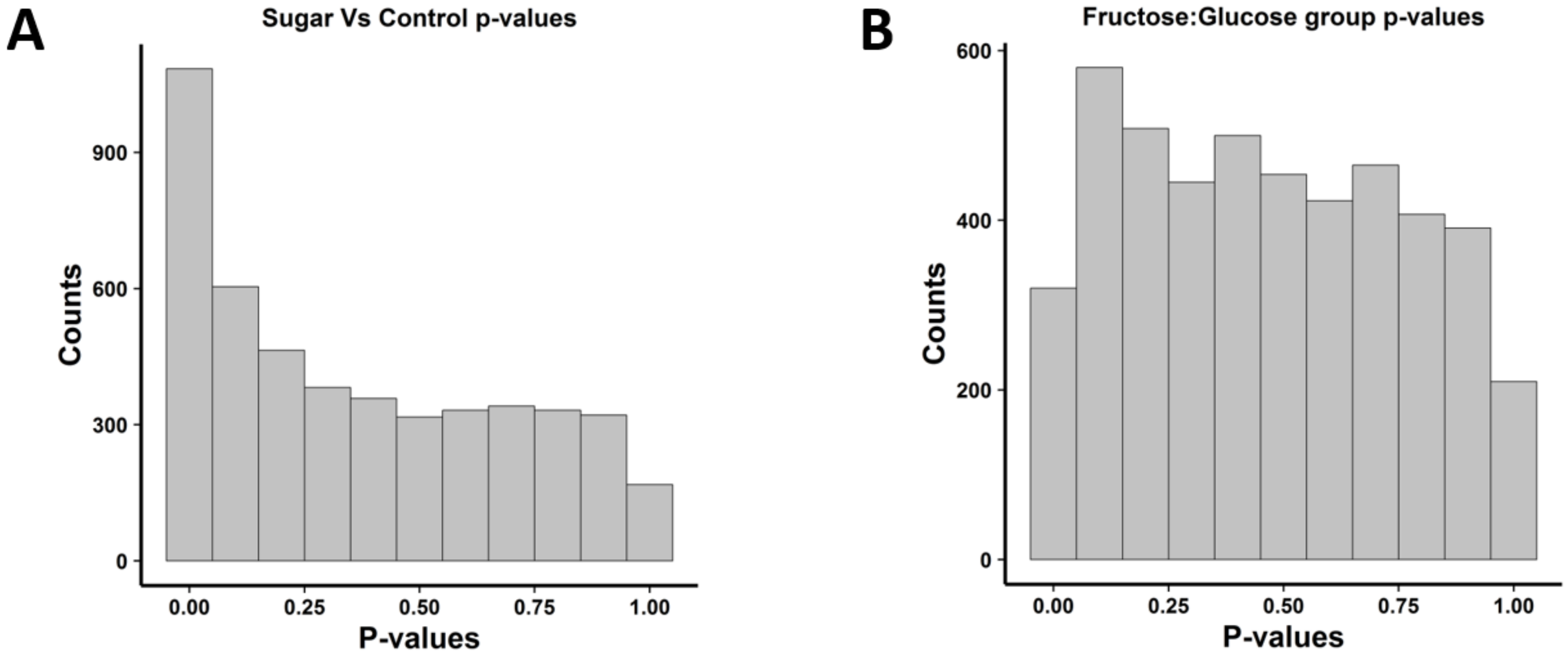

Figure 5.3: Many OTUs differ by sugar versus water consumption while there seems to be little differences in OTU abundances with respect to fructose:glucose concentrations. This is supported by the tendency of the distribution of the *p*-values obtained from dependent variables of a simple linear regression with OTU-level bacteria abundance as an independent variable (Equations 5.1 & 5.2).

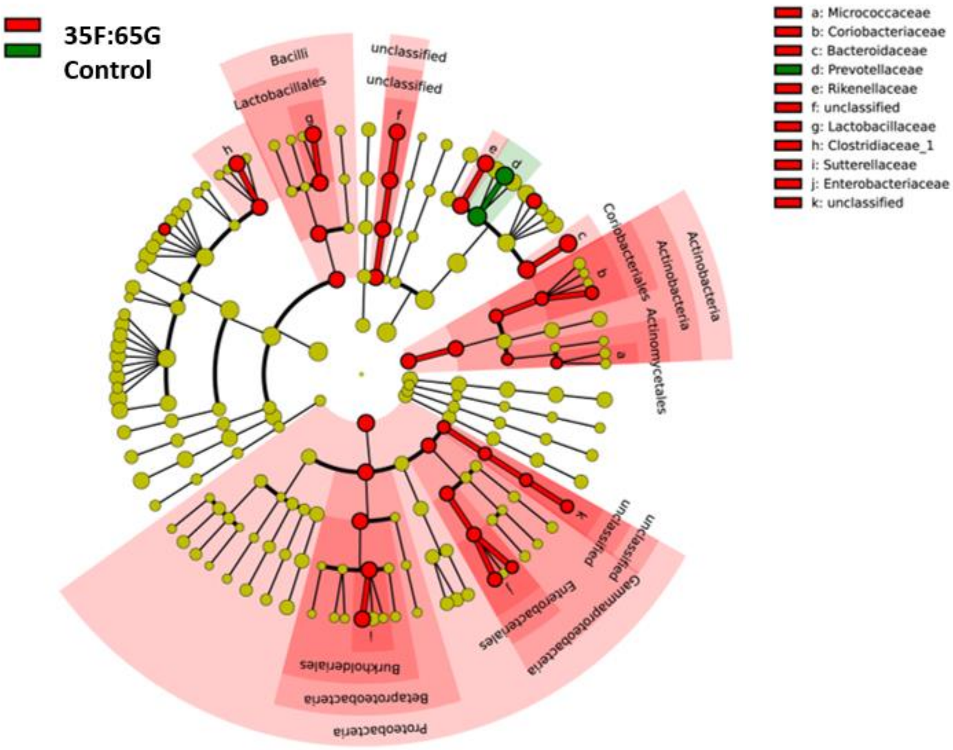

Figure 5.4: LEFsE analysis shows differences between sugar and control at phylogenetic levels.

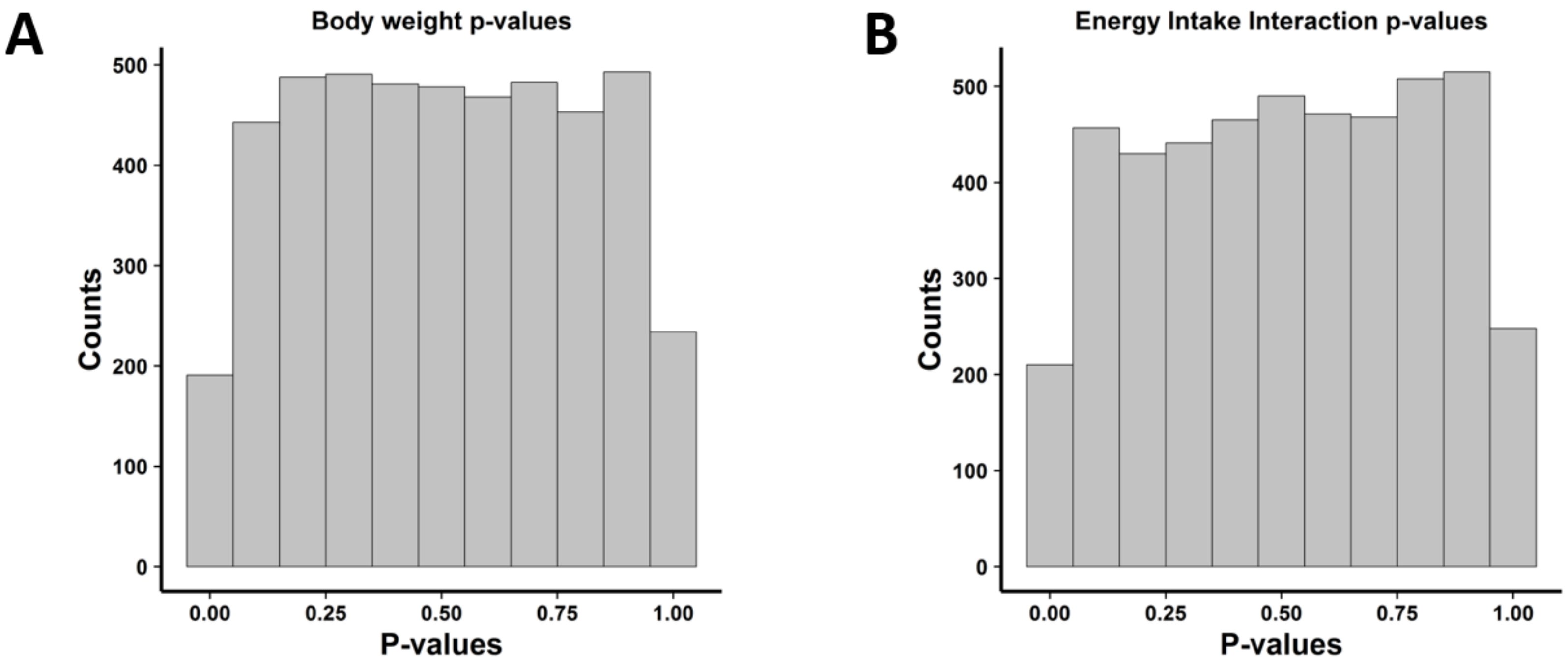

Figure 5.5: Overall, body weight and food intake have little relationship with bacterial abundance in sugar samples. This is supported by the tendency of the distribution of the *p*-values obtained from dependent variables of a simple linear regression with OTU-level bacteria abundance as an independent variable (Equation 5.3).

Table 5.1: Statistics of 16S rRNA sequence reads from fecal samples after various filtering steps.

	Number of Samples	Number of OTUs or Taxa	Number of Sequence Reads	Mean Reads per sample ± SE	Minimum reads per sample	Maximum reads per sample
16S reads generated	42		1,237,456	29,463.24 ± 411.07	21,939	34,034
RDP classified (Phylum Level)	42	11	1,208,210	28,766.9 ± 394.82	21,508	33,392
RDP classified (Class Level)	42	21	1,191,926	28,379.19 ± 393.58	21,191	32,837
RDP classified (Order Level)	42	35	1,187,343	28,270.07 ± 392.22	21,127	32,733
RDP classified (Family Level)	42	84	1,086,649	25,872.6 ± 383.17	19,116	30,283
RDP classified (Genus Level)	42	211	738,112	17,574.1 ± 327.48	12,728	21,924
QIIME OTUs (more than 25% of samples)	42	4,703	918,964	21,880.1 ± 394.77	15,833	28,531

Table 5.2: Numerous bacteria classified at the family are shows significant separation between sugar and control. Shown are those bacteria at the family level whose abundances are significantly different at with a Benjamini-Hochberg corrected *p*-values <0.10. The t-statistic is positive when the mean abundance of the bacteria is higher in sugar samples and negative if higher in control samples.

Name	**Mean abundance in Sugar samples**	**Mean abundance in Control samples**	***t*-statistic**	**Sugar Vs Control *t*-test *p*-value**	**BH corrected *p*-value**
Enterobacteriaceae	1.523928	0.476978	7.081674	2.35E-07	1.67E-05
Carnobacteriaceae	0.900052	0.183798	7.24809	4.03E-07	1.67E-05
Corynebacteriaceae	0.264873	0	4.789857	3.92E-05	0.000813
Bifidobacteriaceae	1.749449	0.60922	4.518762	0.0001312	0.001795
Rikenellaceae	2.970002	2.425188	5.511121	0.0001509	0.001795
Clostridiaceae.1	2.610721	1.580377	4.91767	0.0001514	0.001795
Cryomorphaceae	0.317821	0.057283	4.150955	0.0001962	0.002036
Lactobacillaceae	2.684286	2.26594	4.467189	0.0003783	0.003488
Moraxellaceae	0.420717	0.085992	3.800526	0.0004838	0.004016
Micrococcaceae	0.499164	0.099367	3.866189	0.0005926	0.004472
Bacteroidaceae	3.370912	3.150676	4.101558	0.0007051	0.004877
Prevotellaceae	3.582735	3.804264	-3.52052	0.0020766	0.012311
Coriobacteriaceae	1.842191	1.636179	3.074537	0.0053029	0.029343
Pseudomonadaceae	0.427897	0.190659	2.629361	0.0124471	0.064569
Sutterellaceae	2.565381	2.15203	2.630552	0.0175206	0.085542
Deferribacteraceae	1.278364	1.763033	-2.6485	0.0190084	0.08765

CHAPTER 6: COMPARISON OF MICROBIAL ASSOCIATIONS WITH COLORECTAL ADENOMAS ACROSS MULTIPLE STUDIES

6.1 Abstract

6.1.1 Background. There have been studies that have shown that there is a link between colorectal cancer and the gut microbiome. Reproducibility across studies, however, has yet to be established. In this chapter, we aimed to compare the taxa that are associated with colorectal adenomas status in multiple studies.

6.1.2 Methods. We obtained mucosal biopsy samples from 435 patients (217 with colorectal adenomas and 218 without) in which we extracted 16S rRNA sequence reads and clustered into operational taxonomic units (OTUs). We test for differences in these OTUs due to adenomas status in our dataset plus two previously published datasets (one using mucosal biopsy samples as well and another using fecal samples) from studies in which 16S rRNA sequences and colorectal adenomas status was collected.

6.1.3 Results. We found that there were 59 OTUs that did have a significantly different relative abundance in adenomas versus control samples in our dataset. Taxa were also found to be significant in a re-analysis of previously published datasets. However, the taxa that were different in each dataset did not overlap. That is, no taxa were significantly different in any two of the three datasets. In all three datasets, significantly different taxa tended to be low-abundance taxa.

6.1.4 Discussion. While there are differences in the microbiome of colorectal adenomas subjects, these are driven by low abundant taxa that appear to be irreproducible across

multiple data sets. Larger sample sizes and consistent sampling and sequencing techniques may potentially alleviate this issue.

6.2 Background

While previous studies have shown that there are associations between colorectal cancer and microbial dysbiosis[75], little is known about the causal role the gut microbiome has in colorectal cancer. We know that there is a difference in the richness of the gut microbiota that is associated with adenomas status[17] and also that there is a higher presence of pathogenic bacteria in the gut of subjects who with colorectal adenomas[120]. This chapter examines the relationship between patients who have colorectal adenomas, which is often times an early symptom of colorectal cancer, and the microbial signatures in these patients. In collaboration with Dr. Temitope Keku of the Division of Gastroenterology & Hepatology at the University of North Carolina - Chapel Hill we analyzed 16S sequence data from a total of 435 biopsy samples from 217 patients with colorectal adenomas and 218 people without colorectal adenomas who were used as controls. Utilizing statistical models, we determined which bacteria has a significantly different abundance in subjects with colorectal adenomas versus subjects used as controls. We also compared our results to two previously published datasets: a study published by Sanapareddy et al[17] which demonstrated that on average subjects who have colorectal adenomas had a higher number of taxa in the gut microbial community than those subjects who do not have colorectal adenomas; and a study published by Zackular et al[76] examining the how microbiota could be used as a screening tool for colorectal cancer.

6.3 Methodology

6.3.1 Classify 16S rRNA sequence reads and determine bacteria discriminating case and control samples

To determine the differences in the gut microbial communities of patients in relation to colorectal status, DNA sequences were extracted from mucosal biopsy samples of patients and the 16S rRNA gene was sequenced. 16S sequence reads were separated into separate fastq files per sample and quality-filtered with the default Phred score of 25 using QIIME's "split_libraries.py" script. Paired sequences were merged using FLASh (Fast Length Adjustment of Short reads)[121]. The merged 16S sequence reads were then clustered at a 97% similarity to obtain operational taxonomic units (OTUs) using the UCLUST algorithm[122] in QIIME. We then assigned taxonomy for the OTUs by running a BLAST search for each of the OTU's representative sequence against the GreenGenes database. We only kept OTUs that were present in at least 25% of the samples. To account for differences in raw counts across the samples, the taxonomic classification table was log-normalized (Section 1.6). This resulted in a table with samples as rows and OTUs as columns. We performed multidimensional scaling (MDS) on the resulting table to reduce the table to a matrix in which the columns are eigenvectors that account for some percent of the variation among the samples.

We then ran a non-parametric Wilcoxon rank-sum test[123] for each OTU using the log-normalized abundance of that OTU as the dependent variable and whether the sample comes from a patient with colorectal adenomas (case) or not (control) as the independent variable. We also ran the Wilcoxon rank-sum test using MDS axes instead of OTUs in order to examine the changes in the microbiome as a whole. *P*-values were adjusted for multiple hypothesis testing using the Benjamini and Hochberg method for false discovery rate (FDR) correction.

6.3.2 Collect previously published colorectal adenomas microbial datasets to help validate our findings

In addition to the data provided by Dr. Keku that we are using for this chapter, we also used other gut microbiome studies in which the authors recorded adenomas status for the participants. We used these studies to determine if any microbial associations that we find in our dataset that are different between colorectal adenomas and microbial abundance have a strong enough signal to discriminate between case and control in other datasets as well. To be considered as a possible dataset for validation, the dataset must contain both samples from both subjects with colorectal samples and subjects that will be used as controls. The first dataset that will be used for validation is published by Sanapareddy et al. [17] in which mucosal samples are collected from 71 patients (33 adenomas and 38 control). These samples were collected from subjects as a part of a colonoscopy at the University of North Carolina – Chapel Hill under the guidance of Dr. Keku so it is possible these samples were similar to those in our study because of similar sample origins. This dataset is also similar to our data because the samples are also mucosal biopsies. Our last dataset that is used for validation involves 60 (30 adenomas and 30 control) fecal samples that was published by Zackular et al [76] (Table 6.1). The samples that are used for Zackular et al. dataset was taken from hospitals in Houston, TX, USA; Toronto, ON, CAN; Boston, MA, USA; and Ann Arbor, MI, USA. These were fecal samples that were extracted from subjects who came to the hospitals for a colonoscopy for colorectal cancer screening. We obtained the 16S sequence reads from all these studies and then ran sequences through RDP classifier and QIIME to obtain OTUs in the pipeline described above.

6.4 Results

In an effort to characterize the gut microbiota of patients with colorectal adenomas we sampled 217 patients with colorectal adenomas and also sampled 218 patients who did not have colorectal adenomas as control as part of a screening by colonoscopy. 16S sequence reads were extracted from the samples and these reads were clustered into operational taxonomic units (OTUs) (see Methods).

6.4.1. Determination of taxa that are discriminant for case versus control

A multidimensional scaling (MDS) analysis that there is not much separation between adenomas samples and control samples in our data (Figure 6.1A). However, linear regression models for each of the OTUs with case and control as an independent variable and OTU abundance as the dependent variable (see methods) found 50 out of the 1127 total OTUs that differed in case samples versus control samples at a 10% false discovery rate (Figure 6.1 B).

6.4.2 Microbial associations with colorectal adenomas status was not reproducible in datasets used for validation

In addition to the colorectal adenomas case and control samples that were generated for this study, we sought to determine if OTUs that are associated with case and control were associated with OTU with colorectal adenomas status in other published datasets. We used two outside datasets that are described in Methods section. Although all three datasets aimed to measure gut microbial abundance in patients with and without colorectal adenomas, plotting the first two MDS axes against each other proved that these datasets were strongly different (Figure 6.1) showing a strong batch effect. We would expect these differences given the extraction of the sequences were different (i.e. fecal versus mucosal biopsy; 454 pyrosequencing versus Illumina sequencing; etc.). We noticed that similar to our dataset, in the other two studies there was little separation between the adenomas and control samples based on the first two MDS axes.

Our findings have shown that although there was not much separation by adenomas status in the overall microbial community, there were many closed-reference OTUs that were significantly different due adenomas status in the other two datasets (Figure 6.2). When we plot the average logged abundance of the microbes against the abundance rank of the microbes, we notice that many of the microbes that are significantly different between case and control are ones that have a relatively low abundance. This led us to believe that since these OTUs were low abundant than they are most likely exclusive to their relative dataset which proved to be the case for the most part (Figure 6.3). We collapsed the OTUs to taxa at the genus to the phylum level and found that the three datasets shared a higher number of taxa at these levels. A pairwise comparison between the *p*-values testing the microbial-adenomas associations in the three datasets at all taxonomic levels, however, found little correlation between the microbial-adenomas associations in any of the datasets suggesting scant evidence for a common signal of adenomas across the three datasets (Figure 6.5).

6.5 Discussion

Our study aimed to determine the gut bacteria that were determined to have a different abundance in adenomas and control patients in what is, to our knowledge, the largest study of gut microbiota and adenomas. Despite lack of clear separation of the overall gut microbiota of subjects with colorectal adenomas and those who were used as control, we found that there were over 50 OTUs that had a significant difference in abundance in patients with colorectal adenomas. This finding was in line with previous studies that determined that there were microbes, specifically *Fusobacterium*, that were associated with colorectal cancer[74, 120]. These studies, along with our study, suggests that changes in the gut microbiota could act as an early sign of colorectal cancer as these changes are associated with colorectal adenomas which in itself is a precursor to colorectal cancer. There are also studies showing that *Escherichia coli* was able to induce interleukin-8 (IL-8) which is a pro-inflammatory cytokine[124, 125] allowing for the progression of colorectal adenomas to colorectal cancer.

We not only wanted to determine if microbial associations with adenomas in our study were reproducible in two other published datasets. Surprisingly, while we observed taxa with significant differences between case and control samples in all three datasets, when we collapsed OTUs to taxonomic levels none of these same taxa were significantly different in any two of the datasets.

There are many reasons that could lead to these differences in the datasets including: the type of technology used for the sequencing; the variable region used for the 16S rRNA sequence reads; the type of samples that were sequenced and even the location of the subjects in which the sample is extracted.

Two of the datasets used in this study extracted sequences using Illumina MiSeq. These two datasets shared some of the OTUs although these shared OTUs were not associated with colorectal adenomas status. Not surprisingly, these datasets shared no OTUs with the Sanapareddy et al. dataset that was sequenced using Roche 454 technology.

Illumina MiSeq is a next-generation sequence technique that is able to generate more sequence reads at a fraction of the price of 454 sequencing but with shorter read lengths[126]. This increase number of reads and shorter reads is more likely to create more OTUs which lead to more OTUs being able to be compared across the two sample sets.

Another factor that could possibly drive the differences in the OTUs that are significantly different over the three datasets is the region of the 16S rRNA gene that was used to extract sequence reads. Two of the datasets used V2-V4 region while the other used the V6 region. This makes a difference in the types of OTUs that are obtained because it is possible that one of the regions may evolve at a faster rate than the other regions[127]. This would make it harder to directly compare the OTUs that are generated using the different methods. The region in which there is faster evolution may result in a higher amount of OTUs. Also more tools may be trained using certain variable regions so that could impact accuracy of taxonomic classifications that are assigned to reads[128].

Yet another difference in the datasets is the type of sample that is collected in order to extract the sequences. We have datasets that are using fecal samples and also datasets using mucosal biopsy samples. The datasets from the mucosal biopsy samples would most likely have less bacterial sequences and more sequences that map to human. For this reason, there is at minimum a difference in the amount of bacterial sequences that are generated when using difference sample types. The dataset used for this study also has differences in the location in which the sample subjects live. We have seen this have an effect on the gut microbiome before[93] and it is possible that this is the case for this dataset as well.

This study not only shows us that there are microbes whose abundances are different given adenomas status, but also, surprisingly, that the microbes that are associated with colorectal adenomas does not seem to be the same in multiple datasets. This is discouraging because this can have drastic effect on the reproducibility of the studies that are conducted. This is yet another reason that it is important to create a standard protocol in the field of microbial studies.

Despite the technical reasons that may explain our observed lack of reproducibility, it remains a possibility that different cohorts sampled at different times or in different parts of the country express the gut dysbiosis associated with adenomas in unique ways, and this explains part of the differences between the cohorts that we describe. Further studies will be needed to determine whether the differences we observed between cohorts reflects true biological differences or can be resolved by standardizing collection and sequencing techniques. Resolution of this issue will be crucial if the microbiota are to be used as a diagnostic technique for the presence of colorectal adenomas.

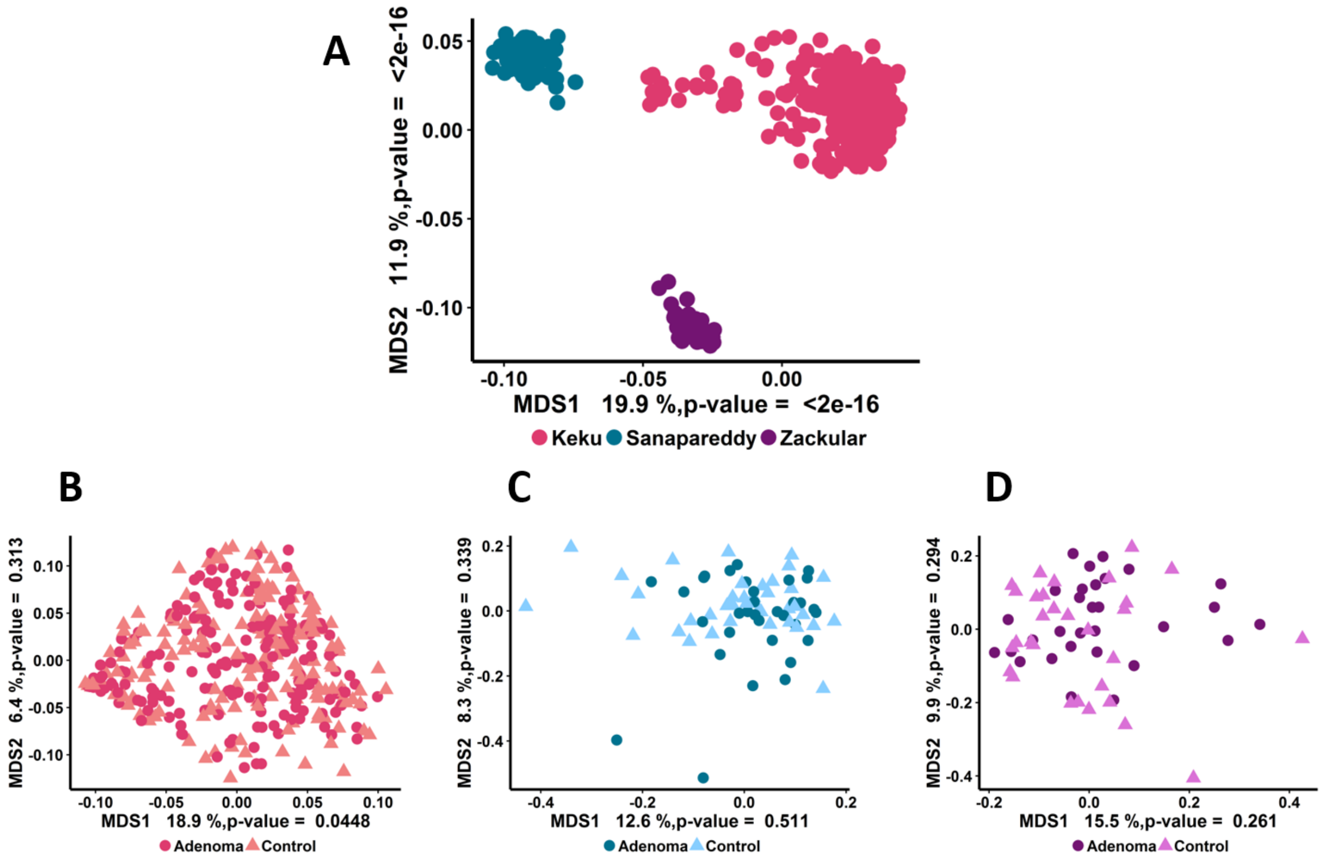

Figure 6.1: Multidimensional scaling (MDS) plot shows that the 16S sequence reads for the three datasets are very different from each other. This difference is bigger than differences in case and control samples within a study (A). (B) our dataset and two previously published data sets: Sanapareddy et al (C) and Zackular et al (D).

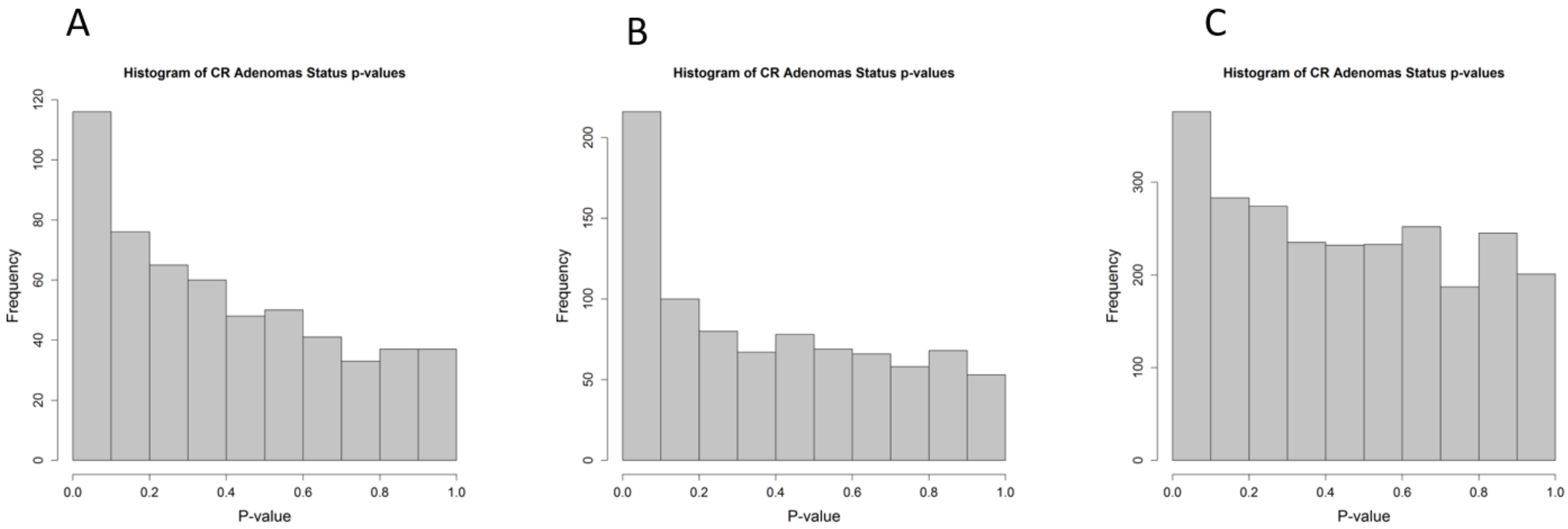

Figure 6.2: Many OTUs discriminate between case and control samples. This is shown when we plot distribution of *p*-values obtained from t-tests with the null hypothesis that there is no significant difference in OTU-level bacteria abundance in case and control samples.

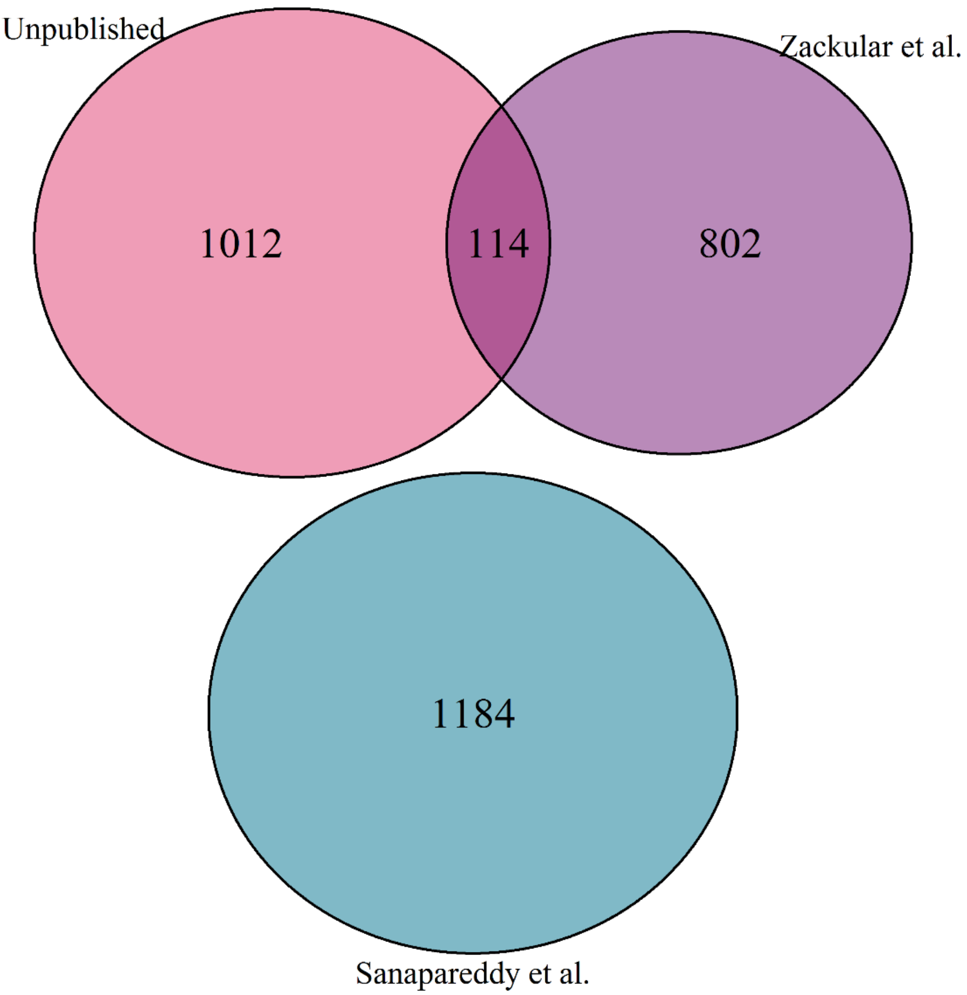

Figure 6.3: Closed-reference OTUs tend to be exclusive to each dataset although our dataset and the Zackular dataset shares a relatively small number of OTUs between each other.

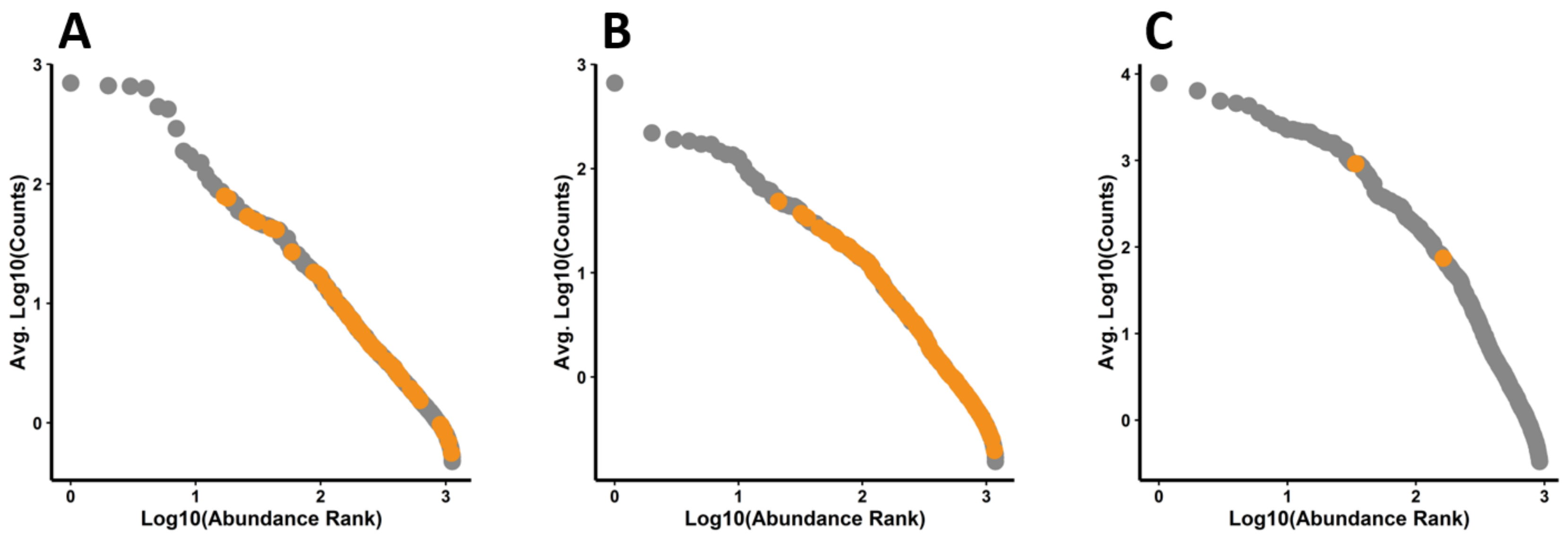

Figure 6.4: OTU Abundance plots colored by significance with orange points representing OTUs with a BH-corrected *p*-value < 0.1. A) Keku unpublished, B) Sanapareddy et al., C) Zackular et. al.

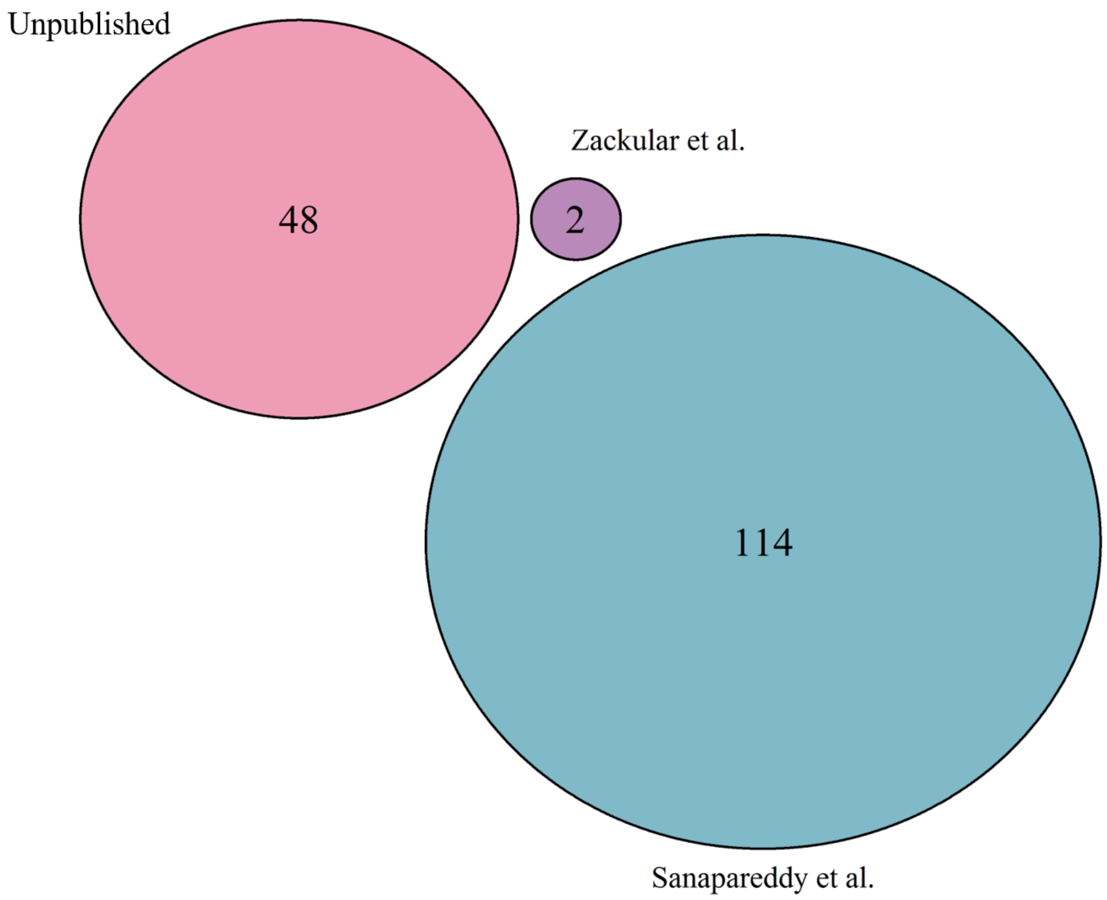

Figure 6.5: OTU Abundance plots colored by significance with orange points representing OTUs with a BH-corrected *p*-value < 0.1. A) Keku unpublished, B) Sanapareddy et al., C) Zackular et. al.

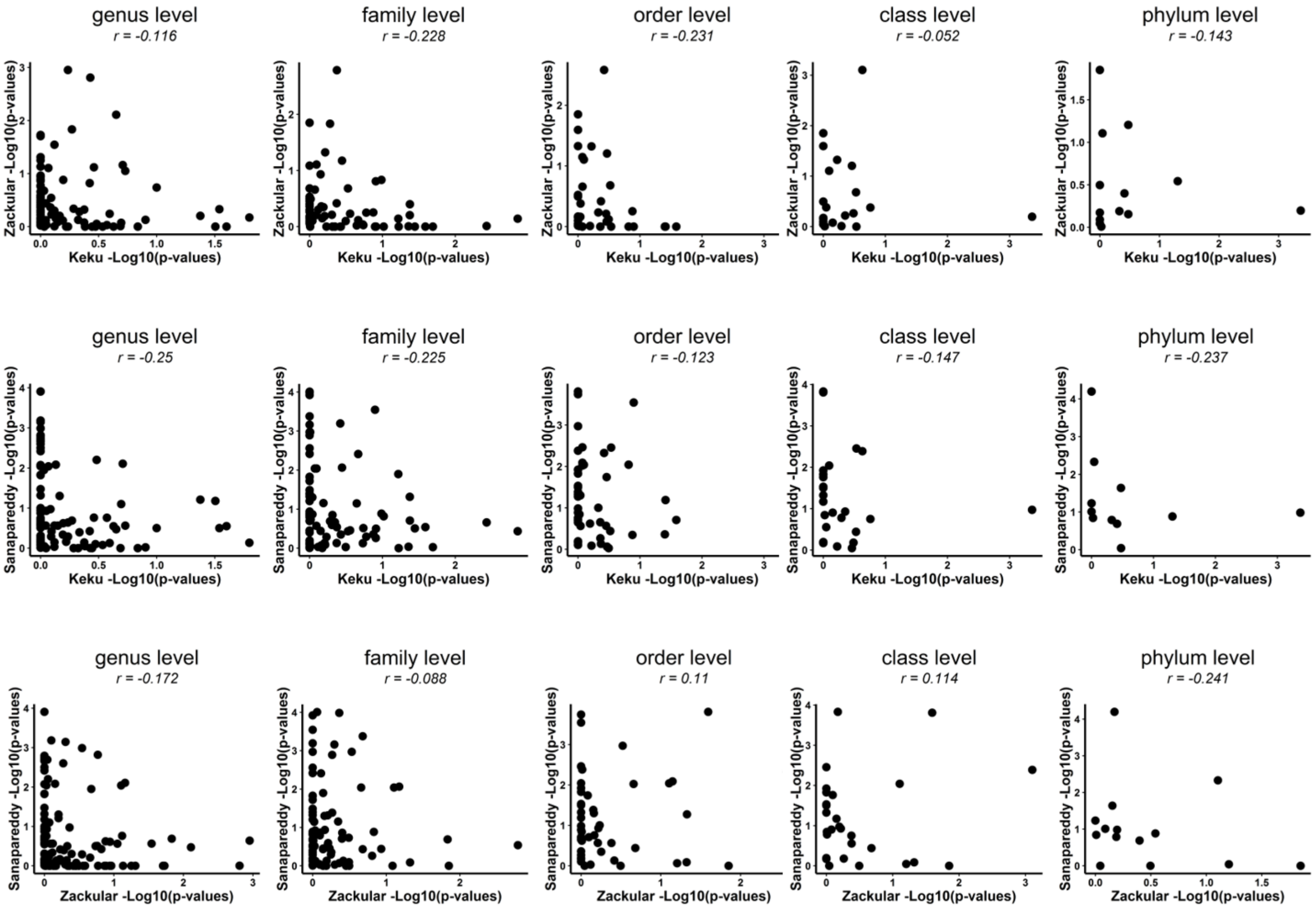

Figure 6.6: Pairwise comparisons of OTU *p*-values in which the means of adenomas and control samples in datasets. We see that at every phylogenetic level there is little correlation between the datasets.

Table 6.1: Description of colorectal adenomas datasets.

Data Set Name	**Keku**	**Sanapareddy**	**Zackular**
Study	Unpublished	Sanapareddy et al., 2012	Zackular et al., 2014
Total Number of Samples	365	71	60
Adenomas	*190*	*33*	*30*
Control	*175*	*38*	*30*
Sample Source	Mucosal Biopsy	Mucosal Biopsy	Fecal
Sequencing Technology	Illumina MiSeq	454	Illumina MiSeq
Study Participants	UNC Colonoscopy patients	UNC Colonoscopy patients	Colonoscopy patients from four locations: Boston, Toronto, Houston and Ann Arbor

CHAPTER 7: SUMMARY

7.1 Conclusion

It has been shown that the trillions of microbes that reside in our bodies are essential for human health. Dysbiosis and lack of diversity in some body sites, is correlated with a number of problems such as obesity, colorectal cancer, and eczema among others. To better understand the relationship the microbiome has on our health, it is imperative to understand who is there (the composition of our microbial communities) and what they are capable of doing (the function of our microbial communities). There are many methods used to quantify the composition and function of microbial communities and this dissertation compares a few of those methods.

In this work, we addressed this by first determining whether swab samples could serve as a replacement method for sampling the microbial community of the human gut. Our findings showed that there was a significant difference between the profiles of the microbial communities of the swab and stool samples. Thus while there would be a loss of information if one sample type was used instead of the other, these samples used in conjunction with one another can provide a more complete view of what is happening in our gut microbiome.

We next calculated how well 16S rRNA sequences, normally used to classify the bacteria in a community, can be used to predict the functions of microbial communities in non-human primates using PICRUSt. This was a vital question because WGS sequences, which are more traditionally used to quantify the functions of community, are expensive in both time and cost. We found that while there was a greater than 85% accuracy of the distribution of the predicted functional profile, these functional predictions were not reliable enough to make statistical inferences. This could be because the majority of fully sequenced bacteria (PICRUSt uses fully sequenced bacteria to help make gene family abundance predictions) are those that are prevalent in humans.

In addition to the analyses focusing on the techniques used to quantify the composition and function of a gut microbial community, we proposed some biological hypotheses in collaboration with researchers at a number of universities. First, we examined the impact of dietary sugar in the gut microbiome of juvenile rats that were fed sugar solutions with varying levels of fructose versus rats that were given water. This work is important because there is a need to better understand the role of the microbiome in metabolizing the fructose that we consume and to see if there is a negative impact of overconsumption of high fructose. I plan to extend this work to examine the impact of sugar on the gut microbiome of adolescents.

Lastly, we sought to find any associations between the gut microbiota and colorectal adenomas and, if present are these associations reproducible across multiple data sets. We found that while there were differences in microbial abundance with respect to adenomas status, these microbes that were different were not reproducible in other datasets. These other datasets had their own set of microbes that were associated with colorectal adenomas status. These differences could be technical differences in the sample type (as we have previously reported with swab and stool samples), sequencing technology and location of the sample source.

. Overall, this dissertation aims to measure the overlap in the techniques used to quantify the microbiome and also aims to apply these techniques to quantify the changes in the gut microbiome associated with high consumption of sugar and with colorectal adenomas. Although the efforts presented in this dissertation will help to gain an understanding of gut microbiome and how it is associated with many external factors such as diet, cage housed, disease state, these are simply associations. There is still much to be understood about the how the microbiome impacts diet and disease and there is a huge gap between our knowledge of what microbes are in the gut and what role does these microbes play. The future of microbiome research will rely on gaining a clearer understanding of the functional pathways that microbes are involved in order to create therapies for conditions that are caused or exacerbated by dysbiosis in microbial communities.

REFERENCES

1. Turnbaugh, P.J., et al., *The human microbiome project.* Nature, 2007. **449**(7164): p. 804-10.

2. Shoemark, D.K. and S.J. Allen, *The Microbiome and Disease: Reviewing the Links between the Oral Microbiome, Aging, and Alzheimer's Disease.* J Alzheimers Dis, 2014.

3. Semova, I., et al., *Microbiota regulate intestinal absorption and metabolism of fatty acids in the zebrafish.* Cell Host Microbe, 2012. **12**(3): p. 277-88.

4. Costello, E.K., et al., *Bacterial community variation in human body habitats across space and time.* Science, 2009. **326**(5960): p. 1694-7.

5. Lozupone, C.A., et al., *Diversity, stability and resilience of the human gut microbiota.* Nature, 2012. **489**(7415): p. 220-30.

6. Human Microbiome Project, C., *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207-14.

7. Kobayashi, T., T. Osaki, and S. Oikawa, *Applying Data Mining to Classify Age by Intestinal Microbiota in 92 Healthy Men Using a Combination of Several Restriction Enzymes for T-RFLP Experiments.* Biosci Microbiota Food Health, 2014. **33**(2): p. 65-78.

8. Deusch, O., et al., *Deep Illumina-based shotgun sequencing reveals dietary effects on the structure and function of the fecal microbiome of growing kittens.* PLoS One, 2014. **9**(7): p. e101021.

9. Claesson, M.J., et al., *Gut microbiota composition correlates with diet and health in the elderly.* Nature, 2012. **488**(7410): p. 178-84.

10. Yatsunenko, T., et al., *Human gut microbiome viewed across age and geography.* Nature, 2012. **486**(7402): p. 222-7.

11. Wu, G.D., *Diet, the Gut Microbiome and the Metabolome in IBD.* Nestle Nutr Inst Workshop Ser, 2014. **79**: p. 73-82.

12. Clarke, S.F., et al., *Exercise and associated dietary extremes impact on gut microbial diversity.* Gut, 2014.

13. Subramanian, S., et al., *Persistent gut microbiota immaturity in malnourished Bangladeshi children.* Nature, 2014. **510**(7505): p. 417-21.

14. Turnbaugh, P.J., et al., *Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome.* Cell Host Microbe, 2008. **3**(4): p. 213-23.

15. Zhang, H., et al., *Dynamics of gut microbiota in autoimmune lupus.* Appl Environ Microbiol, 2014.

16. Highet, A.R., et al., *Gut microbiome in sudden infant death syndrome (SIDS) differs from that in healthy comparison babies and offers an explanation for the risk factor of prone position.* Int J Med Microbiol, 2014. **304**(5-6): p. 735-41.

17. Sanapareddy, N., et al., *Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans.* ISME J, 2012. **6**(10): p. 1858-68.

18. Ley, R.E., et al., *Obesity alters gut microbial ecology.* Proc Natl Acad Sci U S A, 2005. **102**(31): p. 11070-5.

19. Cani, P.D., et al., *Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice.* Diabetes, 2008. **57**(6): p. 1470-81.

20. Schloss, P.D., et al., *The dynamics of a family's gut microbiota reveal variations on a theme.* Microbiome, 2014. **2**: p. 25.

21. Hampton, T.H., et al., *The microbiome in pediatric cystic fibrosis patients: the role of shared environment suggests a window of intervention.* Microbiome, 2014. **2**: p. 14.

22. McCafferty, J., et al., *Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model.* ISME J, 2013. **7**(11): p. 2116-25.

23. Ravel, J. and K.E. Wommack, *All hail reproducibility in microbiome research.* Microbiome, 2014. **2**(1): p. 8.

24. Jun, S.R., et al., *PanFP: pangenome-based functional profiles for microbial communities.* BMC Res Notes, 2015. **8**: p. 479.

25. Langille, M.G., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nat Biotechnol, 2013. **31**(9): p. 814-21.

26. Wu, G.D., et al., *Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags.* BMC Microbiol, 2010. **10**: p. 206.

27. Budding, A.E., et al., *Rectal swabs for analysis of the intestinal microbiota.* PLoS One, 2014. **9**(7): p. e101344.

28. Mokdad, A.H., et al., *Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001.* JAMA, 2003. **289**(1): p. 76-9.

29. Bray, G.A., S.J. Nielsen, and B.M. Popkin, *Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity.* Am J Clin Nutr, 2004. **79**(4): p. 537-43.

30. Walker, R.W., K.A. Dumke, and M.I. Goran, *Fructose content in popular beverages made with and without high-fructose corn syrup.* Nutrition, 2014. **30**(7-8): p. 928-35.

31. Bocarsly, M.E., et al., *High-fructose corn syrup causes characteristics of obesity in rats: increased body weight, body fat and triglyceride levels.* Pharmacol Biochem Behav, 2010. **97**(1): p. 101-6.

32. Zuur, A.F., *Mixed effects models and extensions in ecology with R.* 2009, Springer,: New York ; London. p. 574 p. ill.

33. Yarza, P., et al., *Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences.* Nat Rev Microbiol, 2014. **12**(9): p. 635-45.

34. Fox, G.E., et al., *Classification of methanogenic bacteria by 16S ribosomal RNA characterization.* Proc Natl Acad Sci U S A, 1977. **74**(10): p. 4537-41.

35. Woese, C.R., *Bacterial evolution.* Microbiol Rev, 1987. **51**(2): p. 221-71.

36. Chakravorty, S., et al., *A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.* J Microbiol Methods, 2007. **69**(2): p. 330-9.

37. Thomas, T., J. Gilbert, and F. Meyer, *Metagenomics - a guide from sampling to data analysis.* Microb Inform Exp, 2012. **2**(1): p. 3.

38. Weinstock, G.M., *Genomic approaches to studying the human microbiota.* Nature, 2012. **489**(7415): p. 250-6.

39. Langille, M.G.I., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nat Biotech, 2013. **31**(9): p. 814-821.

40. Gilbert, J.A. and M. Hughes, *Gene expression profiling: metatranscriptomics.* Methods Mol Biol, 2011. **733**: p. 195-205.

41. Gifford, S., B. Satinsky, and M.A. Moran, *Quantitative microbial metatranscriptomics.* Methods Mol Biol, 2014. **1096**: p. 213-29.

42. Huse, S.M., et al., *A core human microbiome as viewed through 16S rRNA sequence clusters.* PLoS One, 2012. **7**(6): p. e34242.

43. Gonzalez, A. and R. Knight, *Advancing analytical algorithms and pipelines for billions of microbial sequences.* Curr Opin Biotechnol, 2012. **23**(1): p. 64-71.

44. Winglee, K.F., Anthony, *Intrinsic association between diet and the gut microbiome: current evidence.* Nutrition and Dietary Supplements, 2015.

45. Gootenberg, D.B. and P.J. Turnbaugh, *Companion animals symposium: humanized animal models of the microbiome.* J Anim Sci, 2011. **89**(5): p. 1531-7.

46. Kostic, A.D.H., M.R.; Garrett, W.S., *Exploring host-microbiota interactions in animal models and humans.* Genes & Development, 2013. **27**(7): p. 701-18.

47. Rawls, J.F., et al., *In vivo imaging and genetic analysis link bacterial motility and symbiosis in the zebrafish gut.* Proc Natl Acad Sci U S A, 2007. **104**(18): p. 7622-7.

48. Brugman, S., et al., *Oxazolone-induced enterocolitis in zebrafish depends on the composition of the intestinal microbiota.* Gastroenterology, 2009. **137**(5): p. 1757-67 e1.

49. Mouse Genome Sequencing, C., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.

50. Ley, R.E., et al., *Evolution of mammals and their gut microbes.* Science, 2008. **320**(5883): p. 1647-51.

51. Nelson, M.C., et al., *Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys.* PLoS One, 2014. **9**(4): p. e94249.

52. Sipos, R., et al., *Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis.* FEMS Microbiol Ecol, 2007. **60**(2): p. 341-50.

53. Lee, C.K., et al., *Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing.* PLoS One, 2012. **7**(9): p. e44224.

54. Pinto, A.J. and L. Raskin, *PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets.* PLoS One, 2012. **7**(8): p. e43093.

55. Hong, S., et al., *Polymerase chain reaction primers miss half of rRNA microbial diversity.* ISME J, 2009. **3**(12): p. 1365-73.

56. Brooks, J.P., et al., *The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies.* BMC Microbiol, 2015. **15**(1): p. 66.

57. Kanagawa, T., *Bias and artifacts in multitemplate polymerase chain reactions (PCR).* J Biosci Bioeng, 2003. **96**(4): p. 317-23.

58. Shinoda, N., et al., *High GC contents of primer 5'-end increases reaction efficiency in polymerase chain reaction.* Nucleosides Nucleotides Nucleic Acids, 2009. **28**(4): p. 324-30.

59. Ahn, J.H., et al., *Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities.* J Microbiol, 2012. **50**(6): p. 1071-4.

60. Hugenholtz, P. and T. Huber, *Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases.* Int J Syst Evol Microbiol, 2003. **53**(Pt 1): p. 289-93.

61. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes.* Nat Methods, 2012. **9**(8): p. 811-4.

62. Paulson, J.N., et al., *Differential abundance analysis for microbial marker-gene surveys.* Nat Methods, 2013. **10**(12): p. 1200-2.

63. Lagier, J.C., et al., *Microbial culturomics: paradigm shift in the human gut microbiome study.* Clin Microbiol Infect, 2012. **18**(12): p. 1185-93.

64. Seng, P., et al., *Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry.* Clin Infect Dis, 2009. **49**(4): p. 543-51.

65. Santiago, A., et al., *Processing faecal samples: a step forward for standards in microbial community analysis.* BMC Microbiol, 2014. **14**: p. 112.

66. Human Microbiome Project, C., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-21.

67. Cardona, S., et al., *Storage conditions of intestinal microbiota matter in metagenomic analysis.* BMC Microbiol, 2012. **12**: p. 158.

68. Huse, S.M., et al., *Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects.* Microbiome, 2014. **2**(1): p. 5.

69. White, J.S., *Straight talk about high-fructose corn syrup: what it is and what it ain't.* Am J Clin Nutr, 2008. **88**(6): p. 1716S-1721S.

70. Hanover, L.M. and J.S. White, *Manufacturing, composition, and applications of fructose.* Am J Clin Nutr, 1993. **58**(5 Suppl): p. 724S-732S.

71. Rippe, J.M., *Fructose, high fructose corn syrup, sucrose and health*. 2014.

72. Hallfrisch, J., *Metabolic effects of dietary fructose.* FASEB J, 1990. **4**(9): p. 2652-60.

73. Siegel, R., C. Desantis, and A. Jemal, *Colorectal cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(2): p. 104-17.

74. Akin, H. and N. Tozun, *Diet, microbiota, and colorectal cancer*. J Clin Gastroenterol, 2014. **48 Suppl 1**: p. S67-9.

75. Sobhani, I., et al., *Microbial dysbiosis in colorectal cancer (CRC) patients.* PLoS One, 2011. **6**(1): p. e16393.

76. Zackular, J.P., et al., *The human gut microbiome as a screening tool for colorectal cancer.* Cancer Prev Res (Phila), 2014. **7**(11): p. 1112-21.

77. Azcarate-Peril, M.A., M. Sikes, and J.M. Bruno-Barcena, *The intestinal microbiota, gastrointestinal environment and colorectal cancer: a putative role for probiotics in prevention of colorectal cancer?* Am J Physiol Gastrointest Liver Physiol, 2011. **301**(3): p. G401-24.

78. Boleij, A. and H. Tjalsma, *Gut bacteria in health and disease: a survey on the interface between intestinal microbiology and colorectal cancer.* Biol Rev Camb Philos Soc, 2012. **87**(3): p. 701-30.

79. Lepage, P., et al., *A metagenomic insight into our gut's microbiome.* Gut, 2013. **62**(1): p. 146-58.

80. Vipperla, K. and S.J. O'Keefe, *The microbiota and its metabolites in colonic mucosal health and cancer risk.* Nutr Clin Pract, 2012. **27**(5): p. 624-35.

81. Karlsson, F.H., et al., *Symptomatic atherosclerosis is associated with an altered gut metagenome.* Nat Commun, 2012. **3**: p. 1245.

82. Devaraj, S., P. Hemarajata, and J. Versalovic, *The human gut microbiome and body metabolism: implications for obesity and diabetes.* Clin Chem, 2013. **59**(4): p. 617-28.

83. Manichanh, C., et al., *Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.* Gut, 2006. **55**(2): p. 205-11.

84. Willing, B.P., et al., *A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes.* Gastroenterology, 2010. **139**(6): p. 1844-1854 e1.

85. Albenberg, L., et al., *Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota.* Gastroenterology, 2014. **147**(5): p. 1055-63 e8.

86. Eckburg, P.B., et al., *Diversity of the human intestinal microbial flora.* Science, 2005. **308**(5728): p. 1635-8.

87. Stearns, J.C., et al., *Bacterial biogeography of the human digestive tract.* Sci Rep, 2011. **1**: p. 170.

88. Espey, M.G., *Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota.* Free Radic Biol Med, 2013. **55**: p. 130-40.

89. Sonnenburg, J.L., L.T. Angenent, and J.I. Gordon, *Getting a grip on things: how do communities of bacterial symbionts become established in our intestine?* Nat Immunol, 2004. **5**(6): p. 569-73.

90. Wu, G.D., et al., *Linking long-term dietary patterns with gut microbial enterotypes.* Science, 2011. **334**(6052): p. 105-8.

91. Araujo-Perez, F., et al., *Differences in microbial signatures between rectal mucosal biopsies and rectal swabs.* Gut Microbes, 2012. **3**(6): p. 530-5.

92. Ukhanova, M., et al., *Gut microbiota correlates with energy gain from dietary fibre and appears to be associated with acute and chronic intestinal diseases.* Clin Microbiol Infect, 2012. **18 Suppl 4**: p. 62-6.

93. Claesson, M.J., et al., *Composition, variability, and temporal stability of the intestinal microbiota of the elderly.* Proc Natl Acad Sci U S A, 2011. **108 Suppl 1**: p. 4586-91.

94. Rajilic-Stojanovic, M., et al., *Long-term monitoring of the human intestinal microbiota composition.* Environ Microbiol, 2012.

95. Arumugam, M., et al., *Enterotypes of the human gut microbiome.* Nature, 2011. **473**(7346): p. 174-80.

96. Koren, O., et al., *A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets.* PLoS Comput Biol, 2013. **9**(1): p. e1002863.

97. Jeffery, I.B., et al., *Categorization of the gut microbiota: enterotypes or gradients?* Nat Rev Microbiol, 2012. **10**(9): p. 591-2.

98. Yong, E., *Gut microbial 'enterotypes' become less clear-cut.* Nature News, 2012.

99. Edwards, T.L., et al., *Genome-wide association study identifies possible genetic risk factors for colorectal adenomas.* Cancer Epidemiol Biomarkers Prev, 2013. **22**(7): p. 1219-26.

100. Hosseini, P., et al., *An efficient annotation and gene-expression derivation tool for Illumina Solexa datasets.* BMC Res Notes, 2010. **3**: p. 183.

101. Aronesty, E., *Comparison of Sequencing Utility Programs.* The Open Bioinformatics Journal, 2013. **7**: p. 1-8.

102. Aronesty, E., *ea-utils: Command-line tools for processing biological sequencing data*. 2011.

103. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nat Methods, 2010. **7**(5): p. 335-6.

104. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.* Appl Environ Microbiol, 2007. **73**(16): p. 5261-7.

105. Maidak, B.L., et al., *The RDP (Ribosomal Database Project).* Nucleic Acids Res, 1997. **25**(1): p. 109-11.

106. Kanehisa, M., et al., *Data, information, knowledge and principle: back to metabolism in KEGG.* Nucleic Acids Res, 2014. **42**(Database issue): p. D199-205.

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

108. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

109. Abubucker, S., et al., *Metabolic reconstruction for metagenomic data and its application to the human microbiome.* PLoS Comput Biol, 2012. **8**(6): p. e1002358.

110. Jari Oksanen, F.G.B., Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner, *vegan: Community Ecology Package*. 2016.

111. Kavanagh, K., et al., *Dietary fructose induces endotoxemia and hepatic injury in calorically controlled primates.* Am J Clin Nutr, 2013. **98**(2): p. 349-57.

112. Balzan, S., et al., *Bacterial translocation: overview of mechanisms and clinical impact.* J Gastroenterol Hepatol, 2007. **22**(4): p. 464-71.

113. DiNicolantonio, J.J., J.H. O'Keefe, and S.C. Lucan, *Added fructose: a principal driver of type 2 diabetes mellitus and its consequences.* Mayo Clin Proc, 2015. **90**(3): p. 372-81.

114. Mitchell, E.L., et al., *Reduced intestinal motility, mucosal barrier function, and inflammation in aged monkeys.* The journal of nutrition, health & aging, 2016: p. 1-8.

115. Ye, Y., *Identification and Quantification of Abundant Species from Pyrosequences of 16S rRNA by Consensus Alignment.* Proceedings (IEEE Int Conf Bioinformatics Biomed), 2011. **2010**: p. 153-157.

116. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

117. Lluch, J., et al., *The Characterization of Novel Tissue Microbiota Using an Optimized 16S Metagenomic Sequencing Pipeline.* PLoS One, 2015. **10**(11): p. e0142334.

118. Paisse, S., et al., *Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing.* Transfusion, 2016. **56**(5): p. 1138-47.

119. Oksanen, J., et al. *vegan: Community Ecology Package*. 2016 6/15/2016; R package version 2.3-4.

120. McCoy, A.N., et al., *Fusobacterium is associated with colorectal adenomas.* PLoS One, 2013. **8**(1): p. e53653.

121. Magoc, T. and S.L. Salzberg, *FLASH: fast length adjustment of short reads to improve genome assemblies.* Bioinformatics, 2011. **27**(21): p. 2957-63.

122. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.

123. Wilcoxon, F., *Individual comparisons of grouped data by ranking methods.* J Econ Entomol, 1946. **39**: p. 269.

124. Martin, H.M., et al., *Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer.* Gastroenterology, 2004. **127**(1): p. 80-93.

125. Hope, M.E., et al., *Sporadic colorectal cancer--role of the commensal microbiota.* FEMS Microbiol Lett, 2005. **244**(1): p. 1-7.

126. Sims, D., et al., *Sequencing depth and coverage: key considerations in genomic analyses.* Nat Rev Genet, 2014. **15**(2): p. 121-32.

127. Noller, H.F. and C.R. Woese, *Secondary structure of 16S ribosomal RNA.* Science, 1981. **212**(4493): p. 403-11.

128. Claesson, M.J., et al., *Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions.* Nucleic Acids Res, 2010. **38**(22): p. e200.

APPENDIX A: CHAPTER 3 SUPPLEMENTAL INFORMATION

Appendix Table 3.1: *P*-values for differences in swab and stool samples for microbes at the genus level.

Bacteria (Genus level)	**Stool Means**	**Swab Means**	**Origin (adj. *p*-value)**	**Individual (adj. *p*-value)**	**Culture Medium**
Peptoniphilus	0.573	2.504	2.00E-44	0.008767758	Anaerobe
Anaerococcus	0.285	2.023	2.01E-42	0.007168463	Anaerobe
Finegoldia	0.307	2.076	5.32E-40	0.174296052	Anaerobe
Anoxybacillus	1.327	2.904	3.12E-35	0.005530238	Aerobe
Thermus	0.208	1.576	1.29E-32	0.110924627	Anaerobe
Geobacillus	1.635	3.119	7.36E-32	0.311352651	Aerobe
Anaerosphaera	0.057	1.281	2.50E-31	0.011836674	Anaerobe
Porphyromonas	0.458	1.800	3.24E-27	2.90E-06	Anaerobe
Acinetobacter	0.067	0.977	1.56E-22	0.129185675	Aerobe
Campylobacter	0.332	1.479	1.76E-22	1.57E-05	Aerobe
Murdochiella	0.120	1.257	3.64E-22	0.002064769	Anaerobe
Prevotella	1.714	2.828	7.93E-22	6.75E-22	Anaerobe
Negativicoccus	0.013	0.747	4.76E-18	0.059020817	Anaerobe
Asaccharobacter	1.372	0.731	1.00E-15	3.79E-17	Anaerobe
Peptostreptococcus	0.183	1.009	4.70E-14	0.000103966	Anaerobe
Corynebacterium	0.104	0.826	8.88E-14	0.315438076	Aerobe
Escherichia_Shigella	0.992	1.828	9.88E-13	2.12E-12	Aerobe
Granulicatella	0.776	0.263	4.19E-12	0.044420738	Aerobe
Varibaculum	0.008	0.466	1.73E-11	0.341347701	Aerobe
Peptococcus	0.154	0.615	3.56E-11	2.50E-06	Anaerobe
Alicyclobacillus	0.028	0.438	1.23E-10	0.258065479	Aerobe
Fusobacterium	0.347	0.991	1.23E-10	5.61E-08	Anaerobe
Fusibacter	1.391	0.912	9.69E-10	1.02E-27	Anaerobe
Mobiluncus	0.054	0.552	1.95E-09	0.159505854	Anaerobe
Gordonibacter	1.411	0.854	4.23E-09	2.78E-09	Anaerobe
Dialister	1.417	2.089	8.83E-09	1.61E-18	Anaerobe
Streptophyta	0.745	0.249	1.02E-08	0.157669247	Unknown
Roseburia	3.425	3.150	1.87E-07	3.45E-10	Anaerobe
Clostridium_XlVa	3.181	3.389	4.79E-07	1.65E-17	Anaerobe
Anaerofustis	0.611	0.272	5.76E-07	0.000648252	Anaerobe
Alistipes	3.346	2.988	1.71E-06	2.88E-19	Anaerobe
Dorea	3.150	2.905	3.52E-06	2.50E-06	Anaerobe
Hallella	0.230	0.679	4.93E-06	5.87E-11	Anaerobe
Brevibacillus	0.066	0.349	1.14E-05	0.625656653	Aerobe
Lachnospiracea_	4.089	3.902	3.54E-05	8.40E-06	Anaerobe

incertae_sedis					
Salmonella	0.142	0.388	9.96E-05	2.11E-11	Aerobe
Aminiphilus	0.309	0.146	0.000492147	2.02E-16	Anaerobe
Clostridium_ sensu_stricto	1.410	1.014	0.000521715	1.33E-16	Anaerobe
Veillonella	1.350	1.056	0.000592687	2.69E-22	Anaerobe
Parvimonas	0.153	0.452	0.000908254	0.003362417	Anaerobe
Acetivibrio	0.515	0.308	0.001523335	1.04E-18	Anaerobe
Ethanoligenens	0.945	0.694	0.001847183	2.36E-16	Anaerobe
Turicibacter	0.833	0.552	0.002147057	1.08E-12	Anaerobe
Phascolarctobacterium	2.673	2.419	0.002215804	1.94E-36	Anaerobe
Coprobacillus	1.307	1.618	0.002411893	3.72E-16	Anaerobe
Lactococcus	1.042	0.731	0.002764843	0.000246714	Aerobe
Actinomyces	1.314	1.004	0.003385329	0.107328641	Aerobe
Enterobacter	0.497	0.779	0.003850286	6.23E-12	Aerobe
Rothia	0.420	0.239	0.004346593	1.05E-08	Aerobe
Bifidobacterium	0.422	0.251	0.004724779	4.91E-10	Anaerobe
Megasphaera	0.317	0.544	0.005613704	1.95E-23	Anaerobe
Sporobacter	0.522	0.354	0.006704599	1.16E-15	Anaerobe
Akkermansia	1.553	1.199	0.00752196	1.50E-14	Anaerobe
Clostridium_IV	3.045	2.863	0.007665329	6.92E-11	Anaerobe
Anaerostipes	1.826	1.586	0.008880383	1.48E-06	Anaerobe
Erysipelotrichaceae_ incertae_sedis	2.148	2.328	0.011109228	4.95E-27	Aerobe
Slackia	0.606	0.444	0.018996542	5.39E-34	Anaerobe
Clostridium_XI	1.894	1.662	0.022150998	1.65E-15	Anaerobe
Allisonella	0.148	0.267	0.031759994	0.001813023	Aerobe
Parasutterella	2.144	2.337	0.063174084	7.71E-44	Anaerobe
Acetanaerobacterium	1.169	1.016	0.068139724	0.008682801	Anaerobe
Subdoligranulum	3.002	2.893	0.074902901	3.82E-37	Anaerobe
Marvinbryantia	0.434	0.566	0.091126708	2.11E-11	Anaerobe
Collinsella	2.231	2.075	0.096709769	2.11E-29	Anaerobe
Gemella	0.697	0.509	0.096709769	0.258065479	Aerobe
Allobaculum	0.521	0.661	0.108278438	1.47E-14	Anaerobe
Bacteroides	4.554	4.506	0.12532248	1.68E-12	Anaerobe
Atopobium	0.348	0.486	0.12602787	0.157669247	Aerobe
Gemmiger	0.954	0.878	0.12602787	1.15E-38	Anaerobe
Lactonifactor	0.596	0.495	0.12602787	0.000505061	Anaerobe
Robinsoniella	0.272	0.373	0.138188163	1.87E-05	Anaerobe
Enterococcus	0.293	0.383	0.1392714	7.89E-27	Aerobe
Anaerotruncus	2.410	2.278	0.157432437	5.13E-15	Anaerobe
Clostridium_XlVb	2.596	2.701	0.157432437	1.29E-14	Anaerobe
Propionibacterium	0.163	0.234	0.162330417	4.08E-05	Anaerobe

Pseudoflavonifractor	2.430	2.517	0.19737728	8.08E-06	Anaerobe
Paraprevotella	1.887	1.788	0.206864475	2.03E-29	Anaerobe
Oribacterium	0.268	0.199	0.219549255	0.000175607	Anaerobe
Sutterella	0.950	1.039	0.219549255	2.53E-37	Anaerobe
Desulfovibrio	0.867	0.760	0.224943627	3.27E-36	Anaerobe
Butyrivibrio	0.825	0.722	0.250002192	2.56E-15	Anaerobe
Succinispira	0.325	0.405	0.256838055	6.04E-36	Anaerobe
Parasporobacterium	0.192	0.140	0.27333161	3.91E-05	Anaerobe
Ruminococcus	2.645	2.545	0.285489331	4.79E-27	Anaerobe
Blautia	3.633	3.569	0.286143306	0.08244119	Anaerobe
Butyricicoccus	2.612	2.561	0.357118294	5.85E-05	Anaerobe
Eubacterium	0.383	0.305	0.363379784	3.60E-06	Anaerobe
Bilophila	1.573	1.638	0.373100237	1.05E-28	Anaerobe
Acidaminococcus	0.409	0.468	0.378145501	2.60E-34	Anaerobe
Klebsiella	0.381	0.307	0.422711268	3.47E-27	Aerobe
Ralstonia	0.427	0.492	0.453455241	0.000212985	Aerobe
Streptococcus	2.736	2.799	0.456861711	1.09E-11	Aerobe
Megamonas	0.226	0.266	0.458810949	4.62E-43	Anaerobe
Hydrogeno-anaerobacterium	0.725	0.659	0.480372208	7.55E-11	Anaerobe
Anaerovorax	1.997	1.949	0.519909747	6.02E-11	Anaerobe
Catenibacterium	0.504	0.438	0.519909747	2.50E-31	Anaerobe
Cloacibacillus	0.273	0.239	0.536501097	7.19E-39	Anaerobe
Clostridium_XVIII	2.603	2.558	0.543957598	2.07E-10	Anaerobe
Faecalibacterium	3.567	3.630	0.565109956	7.12E-36	Anaerobe
Mogibacterium	0.711	0.676	0.586216285	1.68E-14	Anaerobe
Parabacteroides	3.362	3.323	0.586216285	5.08E-38	Anaerobe
Haemophilus	0.539	0.491	0.588769273	2.21E-09	Aerobe
Sphingomonas	0.141	0.168	0.590499898	1.54E-06	Aerobe
Lactobacillus	0.877	0.941	0.633819521	3.66E-05	Anaerobe
Clostridium_XIX	0.377	0.390	0.636197976	5.87E-24	Anaerobe
Syntrophococcus	2.240	2.211	0.636197976	1.40E-24	Anaerobe
Beijerinckia	0.333	0.299	0.645807341	0.00040064	Anaerobe
Holdemania	1.450	1.494	0.667921606	5.70E-07	Anaerobe
Coprococcus	2.787	2.810	0.672182935	2.41E-25	Anaerobe
Odoribacter	2.184	2.162	0.721376895	2.72E-29	Anaerobe
Oscillibacter	3.277	3.294	0.721376895	3.88E-16	Anaerobe
Anaerofilum	0.852	0.820	0.753482449	2.51E-07	Anaerobe
Microbacterium	0.285	0.311	0.753482449	6.54E-18	Aerobe
Howardella	0.747	0.774	0.764992442	2.60E-45	Anaerobe
Hespellia	0.266	0.278	0.812712922	4.58E-05	Anaerobe
Solobacterium	0.494	0.510	0.832128945	2.26E-11	Anaerobe
Lachnobacterium	0.283	0.290	0.898669253	0.096244136	Anaerobe

Eggerthella	1.600	1.602	0.905995774	9.34E-13	Anaerobe
Barnesiella	1.771	1.750	0.942779361	5.69E-45	Anaerobe
Aquabacterium	0.689	0.693	0.980503912	2.10E-43	Aerobe
Flavonifractor	2.667	2.668	0.987162038	2.54E-05	Anaerobe
Butyricimonas	1.488	1.487	0.989886347	6.76E-36	Anaerobe