THE MICROBIOTA MULTIVERSE: FROM GUT TO BRAIN AND BEYOND

by

Matthew C. Brown

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

2019

Approved by:

____
Dr. Anthony Fodor

____
Dr. Cynthia Gibas

____
Dr. Rob Reid

____
Dr. Xinghua Shi

____
Dr. Molly Redmond

ABSTRACT

MATTHEW C. BROWN. The Microbiota Multiverse: From Gut to Brain and Beyond. (Under the direction of DR. ANTHONY A. FODOR)

This work explores the microbiomes of diverse biological contexts with various stakeholders and collaborators. It consists of six objectives: the first two relate to interaction between the gut microbiome and host health, the second two investigate the microbiota-gut-brain axis and the last two move beyond the host to consider the microbiomes of the external environment and controversy in the enterotype hypothesis, a major conceptual framework in human gut microbiome research. The first objective confirms findings in previous research by providing further evidence as to the lack of strong microbial associations seen in the healthy aging of the gastrointestinal tract in a non-human primate. The second objective contributes a negative finding to the discussion of an area of gut health where previous studies claimed to have found associations, but themselves had problems in study design, cohort size, or flawed reporting of statistics. In the third objective, a small anorexia nervosa cohort revealed the persistence of individualized microbiome characteristics even in the course of recovery from severe illness. The findings of a sex-stress interaction in the fourth objective underscore the need for future experiments involving the microbiota-gut-brain axis to use mixed-sex cohorts to yield results suitable for translational research, but also provides further evidence of associations of differentially abundant microbes with stress and anxiety which correspond well with other studies in this field. The evaluation of wastewater processing treatment plants in the fifth objective showed that such facilities are effective in removing pathogens and many genes associated with antibiotic resistance, but may elevate concentrations of antibiotics during the treatment process. The last objective has found that algorithmic methods of determining enterotypes are not robust or consistent subject to dataset choice, normalization strategy and corrections for compositional data. This research is unified through its investigations into what constitutes the proper statistical treatment of metagenomics data, especially in the light of its nature as compositional data, and how this may interact with the creation of meaningful benchmarks and “gold standards” which remain to be discovered or invented for this field in order to support reproducible research.

DEDICATION

This dissertation is dedicated to my family, including cats past and present, and especially my wife, Alicia Tegan Love, who have all shown me love and support throughout this process.

ACKNOWLEDGMENTS

I would like to especially thank my dissertation committee chair, Dr. Anthony Fodor, for many hours of great scientific discussions, encouragement and patience. I am similarly thankful to my dissertation committee members Dr. Cynthia Gibas, Dr. Rob Reid, Dr. Xinghua “Mindy” Shi, and Dr. Molly Redmond. I owe a debt of gratitude to my previous mentors who profoundly influenced my training as a researcher which include Dr. Dennis R. Livesay, Dr. Donald J. Jacobs, and Dr. Jennifer W. Weller. For great scientific conversations and “post-PhD” perspectives, I thank past postdoctoral researchers Dr. Ronald M. Clouse, Dr. Kathryn Winglee and Dr. Madeline R. Galac. I am especially thankful to the collaborators who made aspects of this work possible including Dr. Christian Jobin of the University of Florida, Dr. Kylie Kavanagh of Wake Forest University, Dr. Temitope Keku of the University of North Carolina – Chapel Hill, Dr. Ian Carroll of the University of North Carolina – Chapel Hill, and Dr. Mark Lyte of Iowa State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

16S	16S small subunit ribosomal RNA gene
ALDEx2	ANOVA-like Differential Expression
ALR	Additive Log Ratio
AN	Anorexia Nervosa
ANCOM	Analysis of Composition of Microbiomes
ANOVA	Analysis of Variance
ASM	American Society for Microbiology
ATE	Aeration Tank Effluent
BMI	Body Mass Index
BTBR	Black and Tan Brachyury
CARD	Comprehensive Antibiotic Resistance Database
CF-1	Carworth Farms
CH	Calinski-Harabasz
CLR	Centered Log Ratio
coseq	Co-Expression Analysis of Sequencing Data
CSS	Cumulative Sum Scaling
DESeq	R software for differential gene expression
DHA	Docosahexaenoic Acid
DHMRI	David H. Murdock Research Institute
DMM	Dirichlet Mixture Model
DNA	Deoxyribonucleic Acid
DOWN	Pooled Downstream Sites A and B

DS A	Downstream Site A
DS B	Downstream Site B
edgeR	R software for differential gene expression
ET B1	Enterotype *Bacteroides* Type 1
ET B2	Enterotype *Bacteroides* Type 2
ET P	Enterotype *Prevotella*
ET R	Enterotype *Ruminococcus*
FCE	Final Clarifier Effluent
FDR	False Discovery Rate
HPA	Hypothalamic-Pituitary-Adrenal
HOSP	Hospital Effluent
ILR	Isometric Log Ratio
INF	Influent
JSD	Jenson-Shannon Divergence
LBP-1	Lipid Binding Protein 1
LEfSe	Linear Discriminant Analysis Effect Size
logclr	Logged Centered Log Ratio
logUQ	Log Upper-Quartile
MDS	Multidimensional Scaling
MetaPhlAn2	Metagenomic Phylogenetic Analysis
MT	Microbial Translocation
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health

OTU	Operational Taxonomic Unit
PAM	Partitioning Around Medoids
PC	Pseudo Count
PCE	Primary Clarifier Effluent
PCI	Primary Clarifier Influent
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
PEAR	Paired-End Read Merger
PhILR	Phylogenetic Isometric Log Ratio
QIIME	Quantitative Insights Into Microbial Ecology
qPCR	Quantitative Polymerase Chain Reaction
qRT-PCR	Real-Time Quantitative Reverse Transcription Polymerase Chain Reaction
R	GNU Project Language and Environment for Statistical Computing
RC	Raw Counts in a cell
RDP	Ribosomal Database Classifier
RES	Residential Effluent
rRNA	Ribosomal Ribonucleic Acid
ShortBRED	Short Better Representative Extract Dataset
SPIEC-EASI	Sparse Inverse Covariance estimation for Ecological Association and Statistical Inference
SRA	Sequence Read Archive

SUDD	Symptomatic Uncomplicated Diverticular Disease
TMM	Trimmed Mean of M-Values
TSS	Total Sum Scaling
UniFrac	Unique Fraction Metric
UP	Pooled Upstream Sites A and B
UV	Ultraviolet Disinfected Effluent
V4	Hypervariable Region 4
vegan	R Package for Numerical Ecology
VS	Variance Stabilization
WGS	Whole-Genome Sequencing

CHAPTER 1: OVERVIEW AND OBJECTIVES

1.1 Introduction

Initial estimates of the oft-stated ten-to-one ratio of microbial cells to human cells inside our bodies have been revised to a near one-to-one parity[1]. While not as numerically dominant as initially thought, microbes still offer their host organisms a plethora of functional capabilities not present in the host's own genome[2,3]. In cattle and other ruminants, the microbiota permit the digestion of the cellulose of grasses, and in many plant species it is microbes that facilitate the acquisition of nitrogen. Microbes assist the absorption and production of nutrients in humans as well[4]. Our commensal microbiota also play an important role in training our immune systems to recognize and destroy pathogens[5,6]. In turn, the host provides protection from environmental extremes, and nutrients, among other necessities. It has been demonstrated that the microbiomes of humans have individual signatures and are robust to many forms of perturbation[7,8], which can lead us to pondering the true meaning of "self" when so much of the biological work involved in sustaining a healthy existence is done outside of our own genetic legacies[9,10].

In many biological contexts we are still in the process of conducting a microbial census[11] of "who" is present or absent in a discriminating fashion between the healthy and diseased state[12–14]. There is also the matter of assessing if microbial associations are even strong enough to be considered biologically relevant or experimentally feasible given the costs of sampling from large cohorts, as well as formulating standardized laboratory best practices to better enable comparisons between cohorts and develop modeling strategies to account for such confounders[15,16]. Specific microbes from these metagenomes have been implicated in a diverse array of diseases with many initial discoveries naturally focused on gastrointestinal diseases. However, it has also been appreciated that microbes can have influences distant from the gut. The role of the microbiome in carcinogenesis is a key example of the complex interplay between microbes and host at the systems level of inflammation and at the level of specific molecular mechanisms, which is discussed in our review article "Carcinogenesis and therapeutics: the microbiota perspective"[17] by Tsilimigras, Fodor and Jobin. Initial investigations of microbial relationships with cancers of the gut quickly extended to seemingly unrelated extra-intestinal cancers[17,18]. Interestingly, some microbes are associated with the onset of some cancers[19] while at the same time may offer protection against other cancers[20]. This simultaneous importance of systematic perturbations of the microbiota in response to inflammation and specific molecular mechanisms has been seen in other instances of host-microbiome interactions. There has been increasing support of the role of inflammation in precipitating depression-related mental illnesses, for example, in addition to imbalances of specific neurochemicals[21]. The reach of these organisms extends so far as to influence human behavior and the progression of mental illness, like the well-known example of behavioral changes due to infection by *Toxoplasma gondii*, a parasitic apicomplexan acquired from domestic cats[22]. Other non-infectious mental illnesses, such as anorexia nervosa[23], are now believed to have microbial associations through the microbiota-gut-brain axis[24,25].

Beyond the actions of individual species of microbes in these various biological contexts, some of which act through known molecular mechanisms[26], lies the complexity of microbial community interactions. These communities can act as a biological buffer against invasive pathogens and bring to light the importance of the biological context which most microbes find themselves in determining their contribution to host health or disease. Here dysbiosis, the disruption of normal functions and structure of microbial populations at the systems level, can trigger all manner of negative health outcomes through inflammation and immune system activation[27,28]. Interestingly, the dynamics of these communities can also foster the rapid evolution and dissemination of antibiotic resistance genes through horizontal gene transfer[29,30].

Owing to years of investigations as to the “who” and recent mechanistic explorations of “how”, therapeutic interventions to microbe-driven ailments now complement the punitive actions of antibiotics with probiotics and microbe-supporting prebiotics[31–34]. These interventions are can largely either be restorative, returning the flora to a healthy baseline[35], or supplemental and encouraging a “new normal.” The commensal microbiota even play a role in cancer therapeutics through their metabolism of several prodrugs into active compounds, but they can also interfere with cancer treatments[17]. Understanding and utilizing the microbiome to improve human health is thus a delicate and complicated matter of rethinking intervention strategies against “bad bugs” while supporting our microbial benefactors.

However, investigations of both lab-made communities of probiotics and natural communities present a problem for traditional forms of statistical analysis: most of our current sequencing efforts towards investigating microbiomes can only meaningfully return this census-like data as relative abundances since the actual counts of microbial populations present in these samples have been transformed by the stochastic subsampling processes of sequencing[36]. This reconsideration should not be wholly surprising as most researchers routinely normalize all data from next generation sequencing because the number of sequenced reads from each sample can vary widely for technical replicates of the sample biological sample in the same sequencing reaction which makes working from raw count data alone unreliable and prone to cause errors in making accurate biological inferences. This makes it unreasonable to conclude, for example, that a taxon is considered differentially abundant in sample B versus sample A if the raw sequence counts of that taxon doubles from A to B, when the sequencing depth may vary greatly between the samples which gives the raw counts a very different numerical context. The procedure of using relative abundances from such sequencing data are thus a form of compositional data, data wherein the individual components (microbial taxa) of the mixture are subject to the constraint of summing to a constant, which is usually interpreted to be percentages of a whole[36–38]. Standard statistical analyses will produce artifacts owing to the mathematical differences in this proportional non-Euclidean geometric space[39], and a fuller explanation of these details are presented in section IV of the Background. A summary of working with compositional data in the context of metagenomics and a survey of some approaches to address these statistical concerns appears in our review article “Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges” by Tsilimigras and Fodor[37]. In short, while many solutions exist as publicly available software packages, some biological questions and popular analytic approaches, such as microbial association networks, can only meaningfully be answered and used if additional quantitative measurements are collected during the sequencing experiment[40]. However, a fuller understanding of the impact on biological interpretations and the limitations in the questions that the structure of such data can answer are only beginning to be investigated[36,37,41]. Such field-introspective investigations are crucial in the current atmosphere of the ‘reproducibility crisis’ in science, and the Microbiome Quality Control Consortium has begun to answer some questions related to the laboratory and bioinformatics processes involved in the analysis of microbiome data[42].

1.2 Overall Scope of the Dissertation

This thesis investigates the associations of microbial abundances determined through 16S rRNA sequencing or metagenome whole-genome sequencing (WGS) along with the measurement of various metabolic biomarkers, behavioral assessments, and other metadata covariates. This is done across a variety of biological contexts from model organisms like mice in a controlled laboratory research settings, to more human representative model animals such as non-human primates, to healthy and diseased human cohorts, and to the opposite extreme of samples collected from a wastewater treatment ecosystem. The work is unified by its consideration of the reproducibility of the results of metagenomics studies as seen by replication cohorts, different normalization methods and corrections for compositionality.

1.3 Research Objectives

1.3.1 Objective I: The Influence of Age on Intestinal Microbial Translocation in a Non-Human Primate

Microbial abundances from 16S rRNA gene sequencing of three sample tissue types are assessed alongside biomarkers to investigate age-related differences in the community structure and composition of the microbiota of female vervet monkeys.

1.3.2 Objective II: Microbial Associations with Colonic Diverticulosis

Diverticulosis is a condition wherein the patient has small pouches along the lining of the digestive system called diverticula, but these are usually present without any health problems. However, diverticula can become inflamed or infected leading to diverticulitis through largely unknown etiologies. 16S rRNA sequences of mucosal biopsies from a large cohort of 226 subjects with diverticula and 309 subjects without diverticula were used to assess associations between the microbiota and diverticulosis.

1.3.3 Objective III: Case Study of Daily Changes in Microbial Composition and Diversity in Three Patients with Anorexia Nervosa

Anorexia nervosa has been shown previously to have associations with the intestinal microbial community. Here the microbiomes of a small cohort undergoing hospital-based renourishment is assessed through daily 16S rRNA sampling alongside patient metabolic measures in order to investigate the interactions between the microbiome and recovery from acute anorexia nervosa.

1.3.4 Objective IV: Stress-Sex Interaction Influences the Microbiota-Gut-Brain Axis in a Mouse Model

16S rRNA gene sequencing of a mouse cohort consisting of both males and females is used to investigate differences in microbial abundances according to sex in response to chronic physical stressors. The microbiome data is also assessed for associations with specific behavioral scores from various assessments of anxiety and stress-related behaviors.

1.3.5 Objective V: Microbial Community Composition, Antibiotic Concentrations and Antibiotic Resistance Genes Upstream, Downstream and Within a North Carolina Urban Water System

Microbial abundances and antibiotic concentrations are tracked through two wastewater treatment plants and upstream and downstream of the associated waterways in the Charlotte, North Carolina metropolitan area across multiple time points using whole-genome sequencing in order to study the influence of treatment on relative abundances of pathogenic organisms, genes associated with antibiotic resistance, and concentrations of several common antibiotics.

1.3.6 Objective VI: Probing the Robustness of the Enterotype Hypothesis

Enterotypes are the hypothesized discrete patterns of microbes in the human gut thought to indicative of disease propensity. However, the enterotype hypothesis remains a controversial entity. This aim assesses the robustness of enterotypes through investigating how normalization techniques and compositional corrections influence the clustering methods used in the creation of enterotypes. This will be evaluated on some of the initial datasets used in the formulation of enterotypes in addition to datasets containing much larger human cohorts.

1.4 Expected Significance

This work explores the microbiomes of diverse biological contexts with various stakeholders and collaborators. The first objective confirms findings in previous research by providing further evidence as to the lack of strong microbial associations seen in the healthy aging of the gastrointestinal tract in a non-human primate. The second objective contributes a negative finding to the discussion of an area of gut health where previous studies claimed to have found associations, but those other experiments themselves had problems in study design, cohort size, or flawed reporting of statistics. The case study of a small anorexia nervosa cohort revealed the persistence of a microbiome characteristic of the individual even in the course of recovery from a severe illness. The findings of a sex-stress interaction in the fourth objective underscore the need for future experiments involving the microbiota-gut-brain axis to use mixed-sex cohorts to yield results suitable for translational research, but also provides further evidence of associations of differentially abundant microbes with stress and anxiety which correspond well with other studies in this field. The evaluation of wastewater processing showed that such facilities are effective in removing pathogens and many genes associated with antibiotic resistance, but may elevate concentrations of antibiotics during the treatment process. The last objective finds that algorithmic methods of determining enterotypes are not robust or consistent subject to dataset choice, normalization strategy nor corrections for compositional data, and suggests several revisions to methods to support the communication of biomarkers like enterotypes in a more rigorous manner. Taken as a whole, this research is unified through its investigations into what constitutes the proper statistical treatment of metagenomics data, especially in the light of its nature as compositional data, and how this may interact with the future creation of meaningful benchmarks and “gold standards” which remain to be discovered or invented for this field in order to support reproducible research.

CHAPTER 2: METHODOLOGIES EMPLOYED

2.1 Sequencing and Taxonomic Classification Methods

2.1.1 Amplicon and Whole-Genome Sequencing

Woese famously redefined the “Tree of Life” by his discovery that the 16S rRNA gene could be used as a marker of phylogenetic taxonomy within bacteria[43]. Different variable regions of the gene can be used, and it is well-known that the choice of the 16S variable region being sequenced has the potential to favor some taxa over others[44]. 16S sequencing may not have the strain resolution capabilities of methods based on whole-genome sequencing, but it considered by some researchers to better sample rare taxa[45]. The use of 16S sequencing in these datasets does restrict the microbiome investigated in these experiments towards being primarily bacteria. WGS sequencing was performed on samples collected in aim five in order to characterize the genes and therefore putative mechanisms for conferring antibiotic resistance. It should be noted that WGS approaches to metagenomics also produce compositional data[46]. The specifics of the sequencing experiments are discussed in greater detail in each of the respective sections.

2.1.2 Software for Taxonomic Classification

Taxonomic classification for 16S rRNA gene sequencing is achieved either by mapping using the naïve Bayes-based RDP algorithm[47] using the RDP database[48] or with the QIIME[49] metagenomics analysis suite using the Greengenes database[50]. Results tended to be in high agreement between the two methods when QIIME is set to use a closed-reference to pick OTUs, and this agreement has been observed by others[51]. Unless otherwise specified, only the forward reads were used in classifications as the process of merging reads can drop high-quality forward reads if their lower-quality reverse read fails to pass quality filtering. Taxonomic classifications of WGS sequencing efforts were conducted using MetaPhlAn2[52]. The settings and parameters for each classifier are discussed in each objective. In some objectives, but not all, a software suite being developed by the Fodor lab called BioLockJ towards managing the complexity and reproducibility of bioinformatics and metagenomics pipelines was used.

2.2 Data Transformations and Normalizations

For aims one through four, the taxa counts from these pipelines were log normalized as described in[53]. This is technically considered as a variant of the Total Sum Scaling (TSS) method of normalization. Objective five used relative abundances calculated by the MetaPhlAn2 pipeline for classifying WGS metagenomic data. Objective six considered additional normalization methods or transformations that corrected for compositional data. Many of these normalizations strategies come from Weiss et al.[54] and Pereira et al.[55] and a formula or algorithmic summary for each is presented in Table 2.1. The data transformations that correct for the compositional nature of the data, which should be considered a separated entity from data normalization methods, are covered in section IV of the Background.

Table 2.1: Table of normalizations and their formulas or reference RC = raw counts in a cell, n = number of sequences in a sample, Σx = total number of counts in the table, N = total number of samples, PC = pseudo-count, usually taken to be equal to "1".

Name	**Formula**
Raw Counts	No normalization is performed
Rarefied Relative Abundance	Samples are subsampled (without replacement) to a pre-specified number of counts (usually the smallest number of counts amongst all samples)
Relative Abundance/Naïve Proportion	RC/n

Relative Abundance (Logged)	$\log_{10}\left(\frac{RC}{n} + PC\right)$
Log normalized (Fodor Lab)	$\log_{10}\left(\frac{RC}{n} \times \frac{\sum x}{N} + PC\right)$
Cumulative Sum Scaling (CSS)	[Method from metagenomeSeq used][56]
Trimmed Mean by M-Values (TMM)	[Method from edgeR used][57]
Relative Log Expression (RLE)	[Method from edgeR used]
Log Upper Quartile (logUQ)	Scales counts by the upper quartile value (75th percentile) [Method from edgeR used]
Median Ratio	Performs scaling based on median counts [Method from edgeR used]
DESeq Variance Stabilization	[Method from DESeq used][58]

The most simple normalization to be explored is the naïve/simple proportion or relative abundance, in which the counts of each taxa within a sample is divided by the total number of counts within the sample. This normalization scheme is the default output of several widely used pipelines, including MetaPhlAn2, though in MetaPhlAn2 it takes into account the size of the genomes against which marker sequences are being compared. The log upper-quantile (logUQ) normalization was introduced to scale counts by their upper-quartile/75th percentile such that they are in better agreement with qRT-PCR experimental results. The median ratio normalization uses properties of the median to reduce the influence of potential outliers. Cumulative sum scaling (CSS) uses a sample-determined value in place of the fixed upper-quantile threshold, but the scaling is performed on only the subset of taxa that remain invariant across all samples. Other normalization methods considered, DESeqVS and edgeR-TMM, were developed as parts of bioinformatic tools (DESeq and edgeR respectively) for the analysis differential gene expression data. An underlying assumption for both methods is that the number of genes that are differentially expressed does not constitute a large percentage of the gene expression levels measured, which may not be a reasonable assumption for their usage with metagenomics data. The “VS” in “DESeqVS” stands for variance stabilization. Here OTUs with large count values are scaled such that they do not overwhelm the smaller values present for OTUs. This scaling is done by both row and column. The first scaling factor comes from dividing the count of each OTU by its geometric mean (the product of its mean across all samples). The second divides the count by the median of the scaling factors calculated in the first normalization. These scaled values are then assumed to be taken from a Negative Binomial distribution to evaluate the mean-variance relationship for these abundances. Generalized linear models are then used with these scaled values and distributional assumptions to produce a relationship in which the variance of the OTUs is independent of its mean. The edgeR-TMM (Trimmed Mean by M-Values) normalization also involves steps to limit the influence of OTUs with large count values. This is done by using thresholding to remove those with the highest counts, but the OTUs showing the highest fold changes between the experimental conditions are also removed, which again is telling of the origin of this method in detecting differential gene expression. Then the weighted means of the log-ratios between sample pairs is taken. The final normalization is a function of these scaling factors and the initial library sizes. Note that like other methods that take the logarithms of count data, special accommodations must be made to avoid taking the logarithm of zero values, and this most often takes place by the addition of some pseudo-count value, which is arbitrary determined in many transformations. The normalizations considered could also include rarifying the data in which the counts of all samples are randomly subsampled (without replacement) to a prespecified total number of counts or depth. However, the statistical justification or permissibility of performing rarefication is not without controversy[59]. It should also be mentioned that most of these normalization methods make some assumptions about the structure of the underlying count data, but there remain no firm guidelines in place for selecting an optimal normalization[60]. The ability of such normalization methods to remove the influence of technical artifacts such as sequencing depth on influencing biological conclusions remains an area of active study.

2.3 Statistical Methods

2.3.1 Linear and Mixed Linear Models

Metagenomics datasets tend to be high-dimensional, but underdetermined with the number of different microbial taxa vastly outnumbering the number individual samples collected in all but the largest of studies. This, necessarily, restricts the utility of many analytical tools developed to take advantage of ever-increasing sample sizes in other analytical contexts—microbiome data is “big,” but not in the way the term “big data” is most frequently used with regards to having a large number of records or samples. This data also tends to have a pronounced degree of sparsity, with a large portion of zeros present in the count table. Various forms of linear regression thus form the backbone of our analytical toolkit in determining the differential abundance of microbes between conditions. More complex methods exist and some methods of assessing differential gene expression have been leveraged for use with determining significant differential abundance in metagenomics datasets[56–58]. However, these methods involve making additional assumptions about the distribution of the data, and their true utility in making biologically correct inferences has yet to be rigorously assessed[61,62]. In contrast, the assumptions of linear models are well-known: normality of data, independence, fixed X, and the homogeneity of variance, and nonparametric counterparts free of such assumptions may be used in the cases of larger sample sizes, albeit with decreased power[63–65]. Specialized mixed linear models and other methods for accounting for correlations between samples that would be in violation of the assumptions of the simple linear model can be used for the nesting of individuals into groupings dictated by experimental design or repeated measurements are also used. Such approaches are required to capture the influence of "cage effects" when laboratory animals, especially of concern in the case of mice, share housing with one another to statistically correct for the fact that co-housed animals are expected to have more similar microbiomes than those living in another cage[53].

2.3.2 Nonparametric Models

Nonparametric models are also used as an alternative to safeguard against making biological interpretations of results that could simply be the assumptions of linear modeling carrying through[65]. In this work these mainly comprise the Wilcoxon test (also called the Mann-Whitney U test) for comparing across two conditions, like case and control, the Kruskal-Wallis test for variables representing more than two categories, and Kendall's tau test of correlation between continuous variables. Note that these three tests are unified in their formulation as they all use the U statistic to derive a minimum-variance unbiased estimator, and that the Kruskal-Wallis test is the direct extension of the Wilcoxon test on a categorical variable with more than two levels.

2.3.3 Dimensionality Reduction Through Ordination

Another toolset to assist in both the visual understanding and statistical modeling of high dimensional metagenomics data is that of techniques of dimensionality reduction, also called ordination. In the context of microbiome research, these methods also perform the tasks of assessing the beta-diversity to compare similarities and differences within and across experimental conditions. In this work this is mainly achieved through a constrained version of principal coordinate analysis (PCoA) using the Bray-Curtis dissimilarity by means of the capscale function of the R package vegan[66–68]. The ordination recasts the abundance data into mutually orthogonal vectors explaining the differences between samples. This characterization can also be treated as a response variable for conducting statistical inference. Other ecological measures, like that of alpha-diversity, used in these studies include species richness and Shannon diversity, and these are also calculated using functions from the vegan package.

2.3.4 Multiple Hypothesis Correction Through False Discovery Rate

Important to any statistical analysis, but especially those of high throughput next generation sequencing experiments, is correcting the p-values of the statistical inference to account for the large numbers of hypotheses being tested since a number of p-values will be less than some threshold designating significance by chance. In this work this is done through controlling the Benjamini-Hochberg false discovery rate[69]. Results are said to be significant when FDR-corrected p-values < 0.05, and trends are said to be observed when corrected p-values < 0.10.

2.4 The Compositional Nature of Sequencing Data

2.4.1 Stochastic Origins of Sequencing Data

Sequencing experiments can be thought of a stochastic process sampling randomly from the sequence amplicons to fill up the capacity or number of "slots" in the sequencing chip/machine[36,70]. This stochastic process can produce dramatically different read depths even in technical replicates, and, as mentioned in the normalization section, such data is usually converted to relative proportions or abundances so that taxa can be comparable across samples[60]. Sequencing data is therefore "compositional" by definition in that there is now a constant sum (unity) of the various individual components making up a sample[39].

2.4.2 Familiar Analogy and Spurious Correlation

In everyday life such data appear as the percentages of diet coming from fats, proteins or carbohydrates, for example, as a separate consideration from raw total calories consumed[71]. This numerical constraint introduces negative artifactual correlations that cannot be separated from real biological signals[36]. This phenomenon has been noticed as far back as the late 1800s by Karl Pearson, who observed that three independent and random variables (x, y, z) showing no correlations, two would, upon division by the third (x/z versus y/z), yield substantial spurious correlations[72] [Figure 2.1].

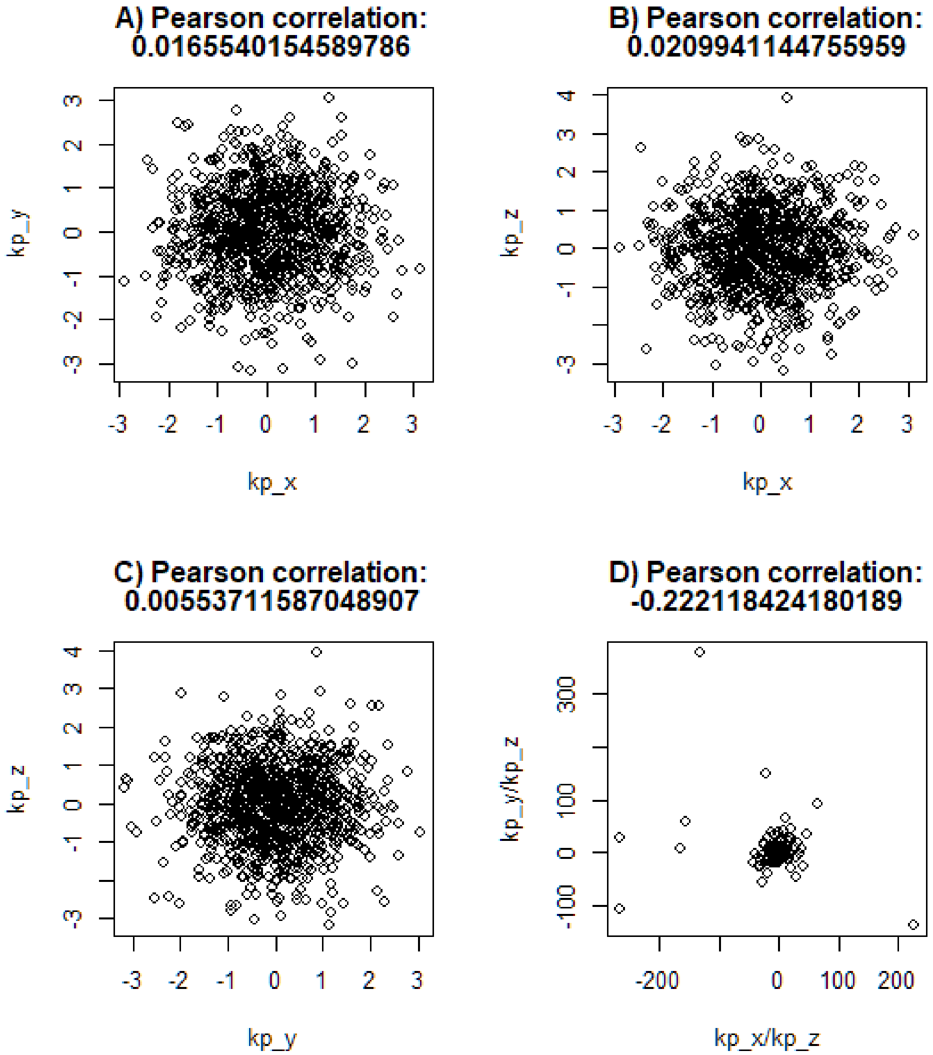

Figure 2.1: Three vectors of 1000 random samples from a standard normal distribution A-C) Random vectors are uncorrelated when considered pairwise D) Upon division by the third vector, spurious correlations occur.

2.4.3 Techniques for Working with Compositional Data

Statistical transformations invented to make compositional data amenable to traditional statistical techniques have their origins in geology where it is used to study mineral compositions of various ores[39,73]. These transformations usually take the form of ratios of logarithms.

This essentially means that the "counts" of biological features of such data (genes or organisms) are not highly correlated to the numerical presence of corresponding sequences, and so it is instead the ratios between features that should be analyzed. This does necessarily represent a fundamental loss of information, and limits the scope of biological questions being asked[37]. For example, questions involving absolute abundance, and potentially even correlations between taxa, require additional evidence beyond such sequencing alone, like qPCR or cell-flow cytometry[74]. Working with such relative data also limits inference. The apparent doubling of the proportion of component A in a three-component mixture (A, B and C) cannot be unambiguously assigned to the absolute increase in the numerical abundance of component A alone as the same pattern would result if B and C decreased in absolute abundance [Figure 2.2].

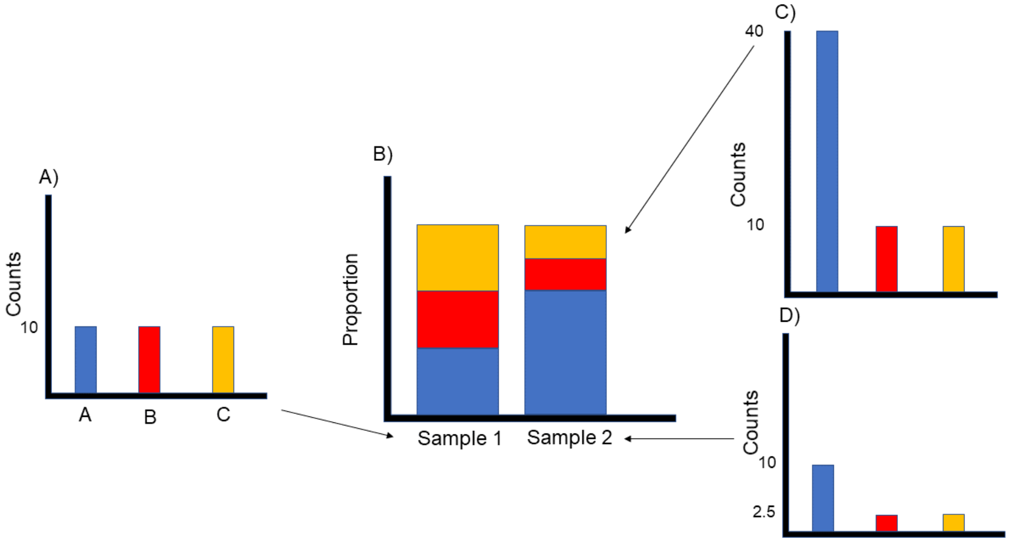

Figure 2.2: Counts versus proportions A) Counts of taxa A, B, C contributing to Sample 1 B) Representation of the proportional doubling of A while B and C remain equal in proportion to each other C) If taxa B and C do not change count relative to panel A) then the absolute count of A would need to quadruple (grow to 40) for a doubling of proportion of taxon A D) If taxon A does not change count relative to panel A) then taxa B and C would need to be reduced to one-quarter (2.5) the original count of 10 for the proportion shown in panel B).

This construction informs an alternative definition of compositional data in that samples C and D are said to have the same “equivalence class” in that when viewed in terms of proportions, they are the same. The ramifications of integrating or ignoring compositional data corrections in the statistical inference of metagenomics is a question that has yet to be fully explored, but it can be broadly considered to be context sensitive to the biology and experiment in question via the sparsity, sample sizes, and microbial interaction network[37,41,75].

2.4.4 The Principal Transformations of Compositional Data Analysis

There are three main statistical transformations for working with compositional data, and all have R or Python packages associated with their use for metagenomic data. The R package ANCOM uses the first developed transform, the additive log ratio (ALR), and normalizes the counts of all taxa to some fixed taxon believed to be invariant across experimental conditions[39,76]. This method is limited in its utility because of the lack of information in selecting the appropriate invariant taxon, as well as the severe consequences for picking one that fails to be invariant, but others have assessed its performance in comparison to normalization methods[60].

$$\mathrm{alr}(\mathrm{x}) = \left[\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \dots, \ln\frac{x_{D-1}}{x_D}\right]$$

Another method, ALDEx2[77], uses the centered log ratio transform where the log ratio is between an individual component and the geometric mean of all components. Zeros in the denominator are replaced by Monte Carlo draws from the Dirichlet distribution which closely resembles the stochastic processes in the sequencing experiment itself. The mapping between the original taxa counts and the transformed values is one-to-one as the dimensionality of the problem remains the same, and so the results tend to be very intuitively interpretable. However, CLR transforms pose limitations as to what downstream statistical tests can be performed with such data as the resulting covariance matrix is singular[78]. It is therefore common practice to transform back to the CLR transformed coordinates owing to the ease of interpretation in this space after performing statistical tests using a transform with more amenable characteristics, but perhaps posing difficulties to interpretation.

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{\mathrm{g}(\mathbf{x})}, \ln \frac{x_2}{\mathrm{g}(\mathbf{x})}, \dots, \ln \frac{x_D}{\mathrm{g}(\mathbf{x})}\right], \mathrm{g}(\mathbf{x}) = \sqrt[D]{x_1 \cdots x_D}$$

The last of the main transformations, the isometric log ratio (ILR)[79], can be seen as an improvement to other methods of transforming compositional data in that it is both orthonormal and metric so that angles and distances between the two spaces correspond as expected. However, the transformation moves the data from a D dimensional space to a D – 1 dimensional space, which poses problems for intuitive interpretation of the resulting transformed data[73]. In the ILR transformation requires the creation of an orthonormal basis (**V**) from the data, and a sequential binary partition is used to iteratively assign the features of the data (taxa) into non-overlapping subgroups called principal balances. There are also many equivalent ways of performing the sequential binary partition required to create the principal balances, which can also lead to interpretation issues. The R package PhILR (**Ph**ylogenetic **ILR**)[80] uses the ILR transformation with the phylogeny guiding the sequential binary partition so that it retains biological interpretability. The Python module gneiss also performs the ILR transformation and uses experimental or environmental covariates, like pH, to guide the sequential binary partitions[81].

$$\mathrm{ilr}_{\mathrm{V}}(\mathbf{x}) = \mathrm{clr}(\mathbf{x}) \cdot \mathbf{V} = \ln(\mathbf{x}) \cdot \mathbf{V}, \text{for matrix } \mathbf{V}\ \mathrm{D} \times (\mathrm{D}-1) \text{such that } \mathbf{V} \cdot \mathbf{V}^{\mathbf{t}} = \boldsymbol{I}_{D-1}$$

The impact of these transformations in the context of determining enterotypes is explored in objective VI, where versions of the transformations are used from the R package coseq[82].

CHAPTER 3: INTESTINAL AGING AND MICROBIAL TRANSLOCATION IN A NON-HUMAN PRIMATE

3.1 Abstract

3.1.1 Background

Aging has the potential to negatively impact gut health through the weakening of mucosal barriers leading to microbial translocation. However, the frequent usage of prescription medicines for other age-related diseases in aging human populations makes studying these interactions in humans difficult, suggesting the use of a non-human primate model.

3.1.2 Methods and Materials

Here 16S sampling of fecal, lumen and mucosal tissue samples from female vervet monkeys classified as either “old” or “young” to investigate changes in microbial associations with age.

3.1.3 Results

No significant associations with age were seen with individual taxa. There are limited differences in the microbial diversity between young and old animals.

3.1.4 Discussion

The limited microbial associations with age agree with other similar studies. However, it should be stressed that this experiment only considers female animals, and whether or not these are consistent with associations in male vervet monkeys remain to be investigated.

3.2 Introduction and Background

The establishment of microbial communities in various human body sites upon birth and infancy is the subject of many ongoing investigations. In many ways, the more challenging questions relate to deciphering the relationships between microbes and aging subject to confounders such as environment and host lifestyle. Advanced age decreases the integrity of the mucosal barrier in the gut leading to more microbial translocation (MT) to extraintestinal sites and the resulting associated inflammation[83,84]. The body's mucosal barriers can be thought of as a kind of internal "skin" separating host from the environment, including food and microbes. This dysfunction of the mucosal barrier of the intestine with age is colloquially known as "leaky gut." The increase in MT is associated with higher levels of endotoxins[85], which have been correlated with insulin resistance leading to metabolic disease[86]. MT is only one of several mechanisms that contribute to immunosenescence, the deterioration of the immune system with age[87,88]. However, there have not been consistent observations of shifts with microbial communities associated with advanced age[89,90].

The use of human cohorts investigating immunosenescence poses many difficulties due to the confounding effects of age, dietary changes and environmental changes. As non-human primates also experience immunosenescence[91] and have a gastrointestinal anatomy similar to that of humans, this makes such animal models attractive alternatives for such studies. These animal models become especially helpful in light of the fact that many non-antibiotic pharmaceuticals impact members of the human microbiome[92], and that pharmaceutical use is prevalent in aged human populations. This aim investigates the changes to the microbial communities associated with age using 16S rRNA gene sequencing within an all-female cohort of vervet monkeys (*Chlorocebus aethiops sabeus*) subjected to a Western diet.

3.3 Attributions

This work is part of a larger study of led by Dr. Kylie Kavanagh of Wake Forest University and the Wake Forest Clarkson Campus/Primate Center. My roles in this project were the processing of sequencing data, and the development and evaluation of statistical modeling pertaining to the microbial analyses present in this work.

3.4 Materials and Methods

3.4.1 Experimental Design

The all-female cohort of vervet monkeys included 9 young and 10 old animals (where "old" denotes an animal more than 18 years old) from the Wake Forest Vervet Research Colony[93]. The animals were fed a standard laboratory diet supplemented with fresh fruits and vegetables. Housing consisted of indoor/outdoor enclosures with *ab libitum* exercise and socialization time. Samples collected for sequencing included fecal, lumen and mucosal scrapings. In this study the collection of different sample types is especially important because the mucosa is involved with host immune defense and is in close contact with the epithelial microbiome, but is often overlooked in favor of the easier to acquire fecal samples[94,95]. Other biomarkers included in this study were qPCR counts from the fecal and mucosal samples, and measurements of the concentration of lipopolysaccharide binding protein 1 (LBP-1), a protein involved in the detection of bacteria and subsequent activation of the immune system[96]. The Wake Forest University Institutional Animal Care and Use Committee gave approval of the protocol in following with the recommendations in the Guide for Care and Use of Laboratory Animals. This included participation in guidelines established by the USDA Animal Welfare Act and Animal Welfare Regulations.

3.4.2 Sequencing and Sequence Processing

The DNA was sequenced using Illumina MiSeq PE250 by the HudsonAlpha Institute for Biotechnology (Huntsville, AL). Demultiplexed sequences are available at the SRA via accession SRP139357. Taxonomic classification was performed using the forward sequencing reads by the RDP algorithm (version 2.6) using the RDP database[48] with default settings and a confidence threshold of 80%[47]. QIIME version 1.8[49] was used alongside for closed-reference taxonomic classification using default parameters against the Greengenes[50] database version 13.5 at a 97% OTU identity. These count tables were log normalized as described previously[53].

3.4.3 Statistical Modeling

Linear modeling was done on each of the three tissue types separately and mixed linear models were created using the lme function of the nlme package in R for all three sample types nested in the source animal as a random effect accounting for the individual signature of the animal. Initial modeling approaches attempted to account for the shared housing of different animals, but the small overall number of subjects and the numerical instabilities inherent to these models ultimately made such an approach unfeasible (data not shown). The response variables included the log normalized abundance of each taxa across the phylum, class, order, family and genus taxonomic levels, the Shannon diversity index for each taxonomic level, and the first several principal coordinates from the beta-diversity ordination. Explanatory variables include age and age-group. Additional models included the Wilcoxon test using age-group as the explanatory variable. The within-sample alpha-diversity was calculated using the Shannon diversity function within vegan. The between-sample beta-diversity was calculated using the capscale function within vegan using the Bray-Curtis dissimilarity. Statistical tests were corrected for multiple hypotheses via the Benjamini-Hochberg FDR correction to a threshold of 0.05. The relevant R scripts are available through GitHub at https://github.com/mcbtBINF/IntestinalAging/

3.5 Results

3.5.1 Microbial Community Assessment Through Beta-Diversity

Methods of assessing the microbiome at the community level demonstrated differences with respect to the sample type, but presented limited differences with respect to age or age-group. Ordination by principal coordinate analysis (PCoA) shows no clear separation by age or age group across all taxonomic levels considered (phylum, class, order, family and genus), but it does show the expected clustering by sample type [Figure 3.1].

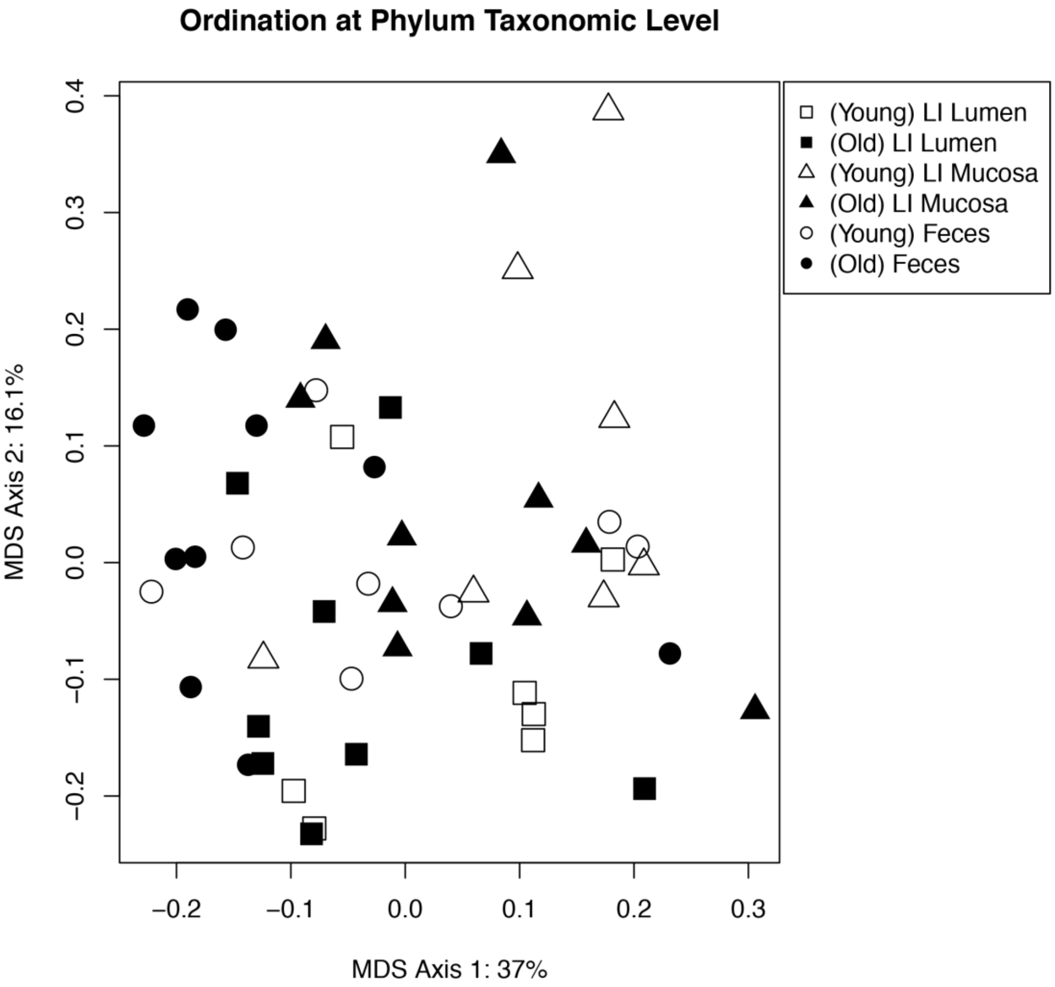

Figure 3.1: Ordination separates on sample type and not on age grouping or numeric age. Figure from Wilson *et al*. 2018[97] (Author's own work).

3.5.2 Intra-Sample Alpha-Diversity as Seen with Shannon Diversity

Similarly, the Shannon diversity of the mucosal samples is significantly less than that of the other sample types at the phylum taxonomic level [Figure 3.2].

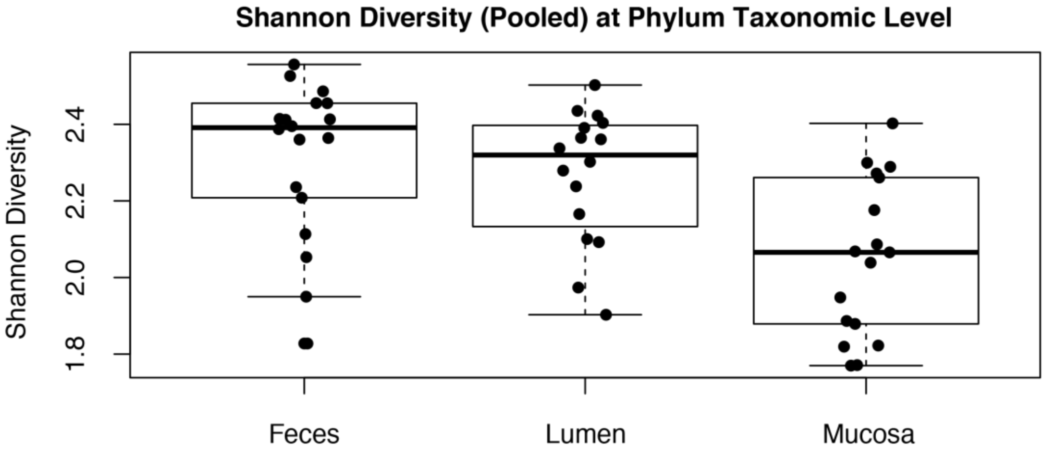

Figure 3.2: The mixed linear model pooling sample types showed significant differences in the Shannon diversity at phylum (depicted) p-value < 0.001, class: 0.001, order: 0.021, family: 0.148, and genus: < 0.001. Figure from Wilson *et al*. 2018[97] (Author's own work).

At various taxonomic levels, the Shannon diversity index was significantly different between the young and old animals [Figure 3.3].

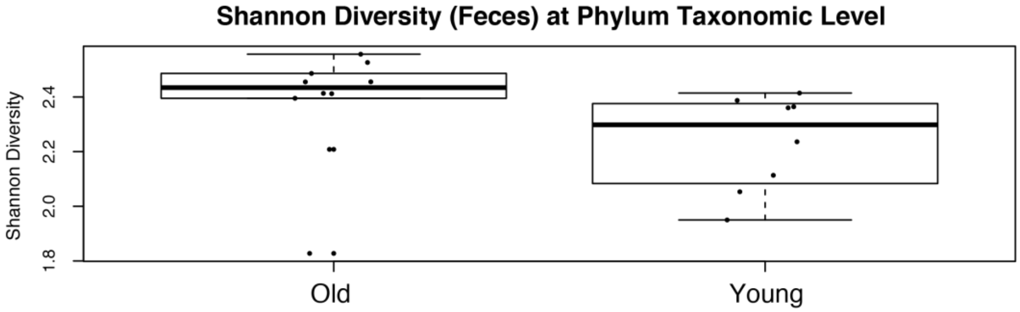

Figure 3.3: The Mann-Whitney-Wilcoxon test showed significant differences between young and old animals at various taxonomic levels (phylum depicted, p-value = 0.033; class: 0.041, order: 0.051, family: 0.016, and genus: 0.286) for the fecal samples, but not for the other sample types. Figure from Wilson *et al*. 2018[97] (Author's own work).

3.5.3 Modeling of Associations Between Specific Taxa and Age

An analysis of individual taxa across the different taxonomic levels did not reveal strong evidence for associations with the animal's age or age-group [Figure 3.4].

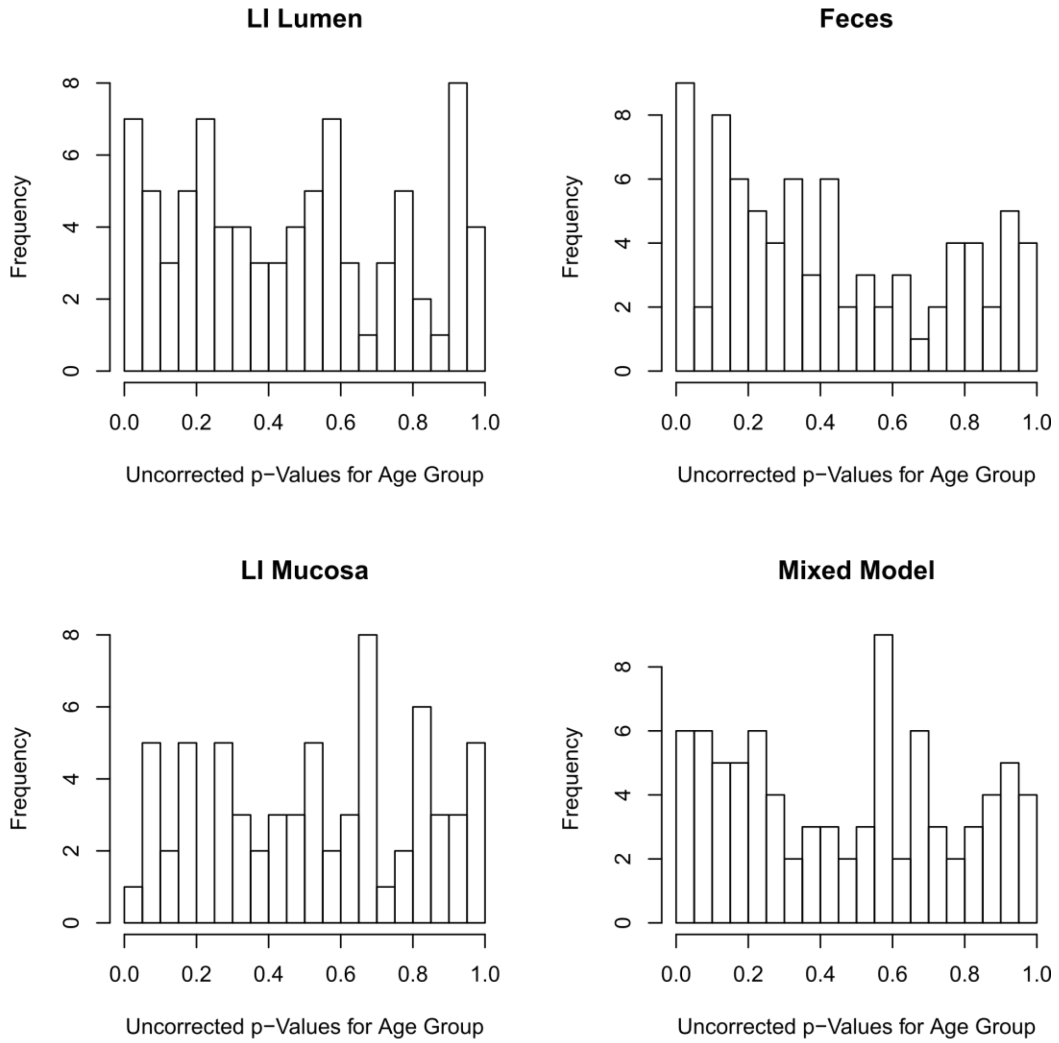

Figure 3.4: Histograms of uncorrected p-values at the genus level for each tissue type Bottom right panel is for the mixed linear model pooling all tissue types grouped by animal of origin. These patterns suggest that there are no strong associations between age-group and specific genera. Figure from Wilson *et al*. 2018[97] (Author's own work).

In addition to evaluating associations between age or age-group with the microbiome, associations between other biomarkers measured in the study and the microbiome were also evaluated. However, there was similarly a lack of strong associations between the microbiome and the biomarkers.

3.6 Discussion

Overall, there is little evidence for the association of individual taxa with age, but this result reproduces what has been seen in previous studies in this host model[90] and helps to establish the robustness of these results. It is therefore concluded that the loss of mucosal barrier function does not have its origins in the microbiome. The discovery of lower diversity in the mucosal tissues is attributed to with function of the mucosa in favoring the survival of select taxa to reduce disease development[98,99].

Other experimental methods used within the study suggested that microbial translocation was occurring at a higher rate in older monkeys. qPCR data also yielded visual trends suggestive of bacterial overgrowth in the mucosal tissues of older monkeys, but these were not statistically significant. These data suggest that aging may lead to lower control over microbial selectivity and colonization at the mucosal surface which is permissive for MT. It should be reiterated that these findings are for a female-only cohort, and results using male animals may differ.

3.7 Communication of Results

These findings have been published in *Scientific Reports* under the title “Greater Microbial Translocation and Vulnerability to Metabolic Disease in Healthy Aged Female Monkeys” by Wilson et al[97].

CHAPTER 4 – MICROBIAL ASSOCIATIONS WITH COLONIC DIVERTICULOSIS

4.1 Abstract

4.1.1 Background

The potential microbial associations with the development of diverticula have been previously explored, but not in a large cohort such as this one with appropriate statistical methodology.

4.1.2 Methods and Materials

16S sequences were generated from mucosal biopsies of 226 subjects with diverticula and 309 subjects without diverticula. These microbial communities were assessed for associations with multiple patient demographic data as well as the presence or absence of diverticula and their count if present.

4.1.3 Results

There are limited microbial associations with the presence/absence of diverticula and with the number of diverticula if present.

4.1.4 Discussion

The null findings of this study conflict with reports from previous literature on the subject, but those studies had small cohort sizes and errors in their reporting of statistics. However, as with any negative result, alternate means of assessing the two experimental groups may result in differences between the conditions.

4.2 Introduction and Background

Diverticulosis, also called symptomatic uncomplicated diverticular disease (SUDD), is the normally asymptomatic condition of having millimeter-scale pouches called diverticula along the intestinal tract[100], and it is prevalent within the aged US population[101]. When these pouches become inflamed or infected, diverticulitis, a condition is associated with numerous negative health outcomes, is the result. This study assessed whether the microbiomes were different between healthy control subjects and otherwise asymptomatic patients with diverticulosis. Associations between the microbiome and the number of diverticula were also investigated. Understanding such relationships could help to prevent the transition from asymptomatic diverticulosis to complications from diverticular diseases.

4.3 Attributions

This work is part of a larger study of led by Dr. Temitope Keku at UNC. My primary role in this project was the verification and independent implementation and extension of statistical models initially explored by Dr. Roshonda Barner-Jones and Dr. Fodor, and to ensure correctness and robustness of results as well as pursuing analytical approaches suggested by reviewers, such as LefSe[102] and microbial association network analysis.

4.4 Materials and Methods

4.4.1 Experimental Design

The study consisted of 226 subjects with diverticula and 309 subjects without diverticula. The tissue samples for the 16S rRNA gene sequences were mucosal biopsies rather than fecal samples owing to the adherent nature of such bacteria and convenience surrounding the logistics of the colonoscopy. These samples and other pertinent patient data were collected by specialists in endoscopy at the Meadowmont Ambulatory Endoscopy Center, UNC Hospitals, Chapel Hill, NC. The subjects provided informed consent according to the appropriate guidelines and the study was approved by the UNC Office of Human Research Ethics.

4.4.2 Sequencing and Taxonomic Assessment

DNA was extracted and the V2 region of the 16S gene was amplified using PCR primers. The study design also included the sequencing of a positive control using known bacteria. Raw sequences are available from the NCBI's SRA repository via SUB3467354 within BioProject PRJNA429136.

Taxonomy was assigned to these sequencing using both the RDP (version 2.10.1) classification algorithm on the RDP database at the 50% confidence level as well as the QIIME 1.91 pipeline using the Greengenes database. Samples with less than 1000 assignments were excluded from downstream analysis, but more than 90% of all samples exceeded this threshold.

4.4.3 Statistical Modeling

Statistical modeling included univariate linear regression where the metadata (case/control, sex, race, diverticula count or waist circumference) served as the explanatory variable with the log-normalized taxon abundance, Shannon diversity or MDS axis serving as the dependent variable. Nonparametric models were similarly constructed using the Wilcoxon, Kruskal-Wallis or Kendall test as appropriate.

4.4.4 Additional Analytical Methods Used

Reviewers to the initial manuscript submission suggested pursuing a linear discriminant analysis via the LEfSe tool, as well as looking for differences in the microbial association networks between patients with and without diverticula. SPIEC-EASI[103], a microbial association network analysis package that corrects for data compositionality was used to investigate the microbial association networks.

4.5 Results

4.5.1 Microbial Community Comparisons Through Beta-Diversity

Case and control were not significantly different with respect to the largest PCoA axes [Figure 4.1].

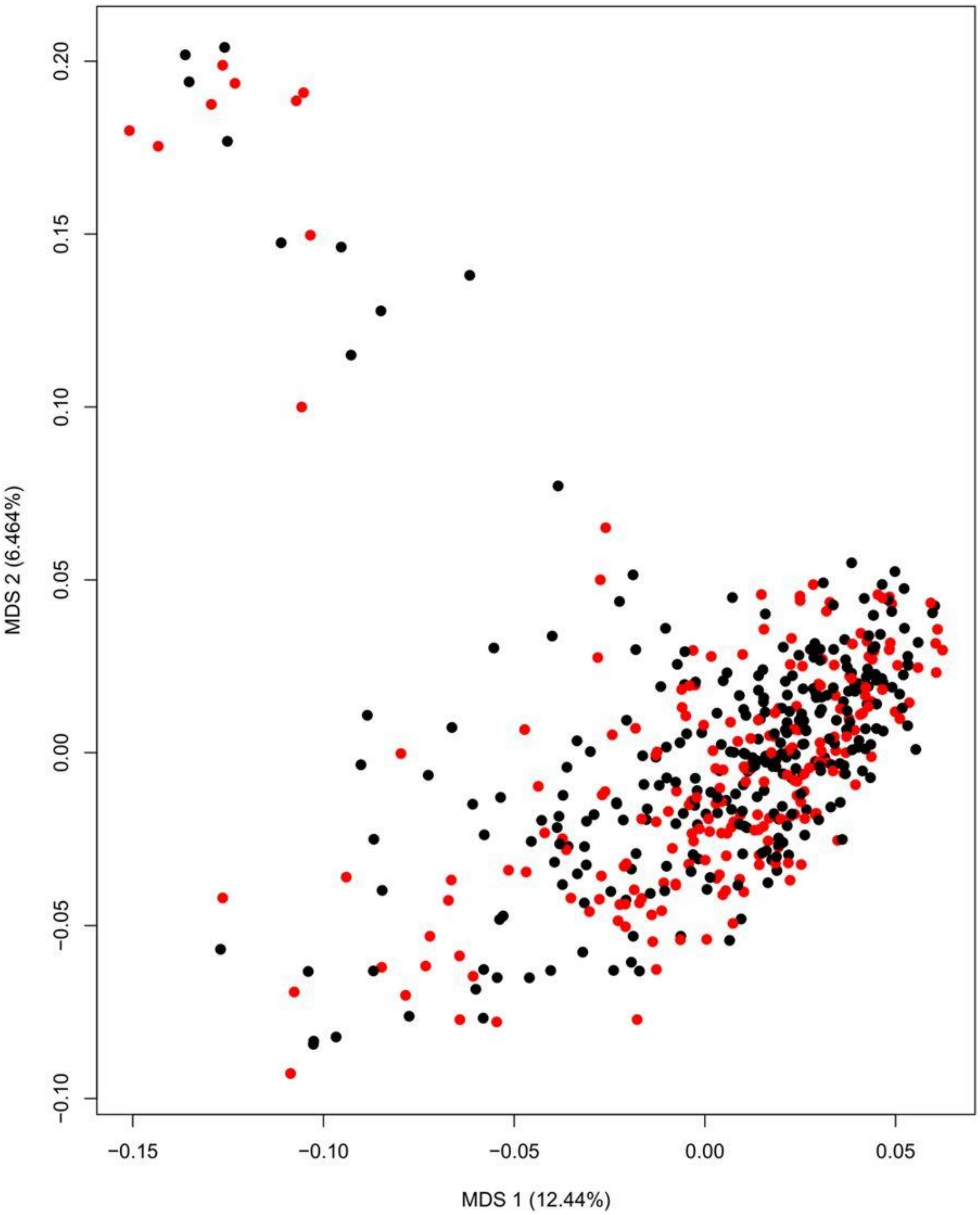

Figure 4.1: Case (red) and control (black) showed no significant associations with largest MDS axes when using the unpaired Wilcoxon testFigure from Jones *et al*. 2018[104] (Author's own work).

The assessment of the alpha-diversity of case and control samples yielded weakly significant differences between case and control [Figure 4.2].

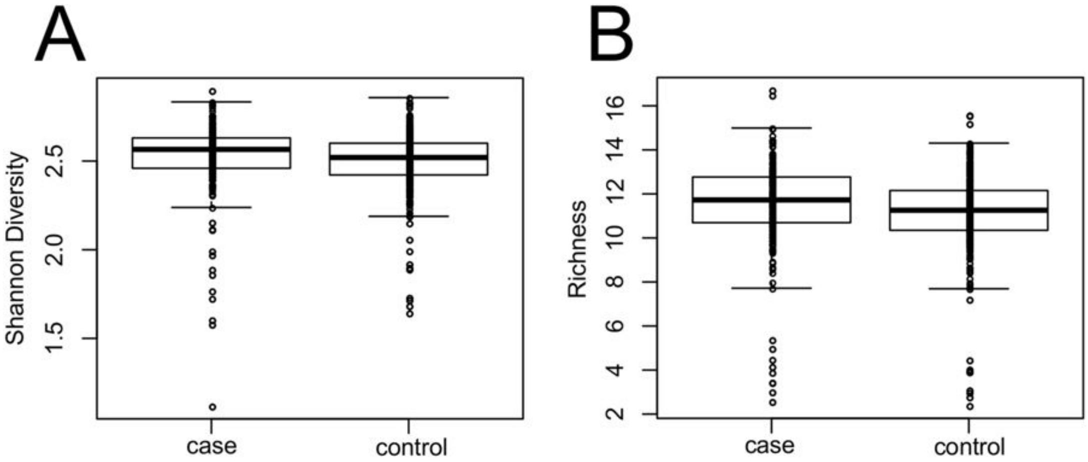

Figure 4.2: Wilcoxon test of case versus control samples at the class taxonomic level yields p-values = 0.011 and 0.012, respectively. An analysis of the same data using linear models produced r-squared values <1%. Figure from Jones *et al*. 2018[104] (Author's own figure).

In terms of differential abundance of specific taxa, the principal findings of this investigation were that only the phylum Proteobacteria and its family Comamonadaceae were marginally significantly differentially abundant between patients with and without diverticula (corrected p-values of 0.038 and 0.035, respectively), but with effect sizes on the order of 2% [Figure 4.3].

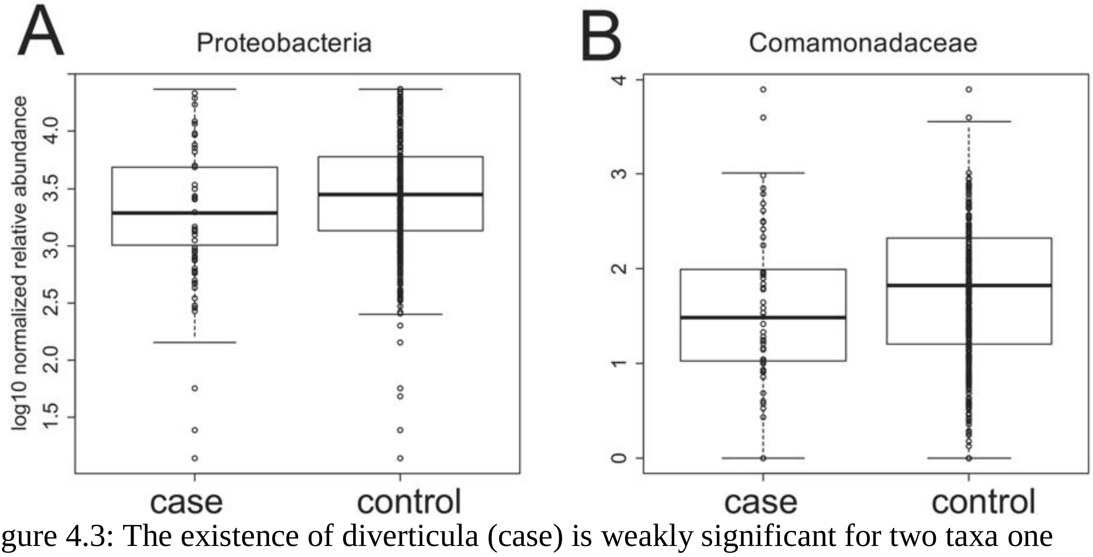

Figure 4.3: The existence of diverticula (case) is weakly significant for two taxa one (Proteobacteria) at the phylum level (p-value = 0.038) and one (Comamonadaceae) at the family level (p-value = 0.035) when assessed with a Wilcoxon test. A similar analysis using a linear model yielded effect sizes on the order of 2%. Figure from Jones *et al*. 2018[104] (Author's own work).

Similarly, limited associations were seen between the diverticula counts and their location, proximal or distal, within the sigmoid colon (data not shown). Results from the RDP and QIIME pipelines were in good agreement [Figure 4.4].

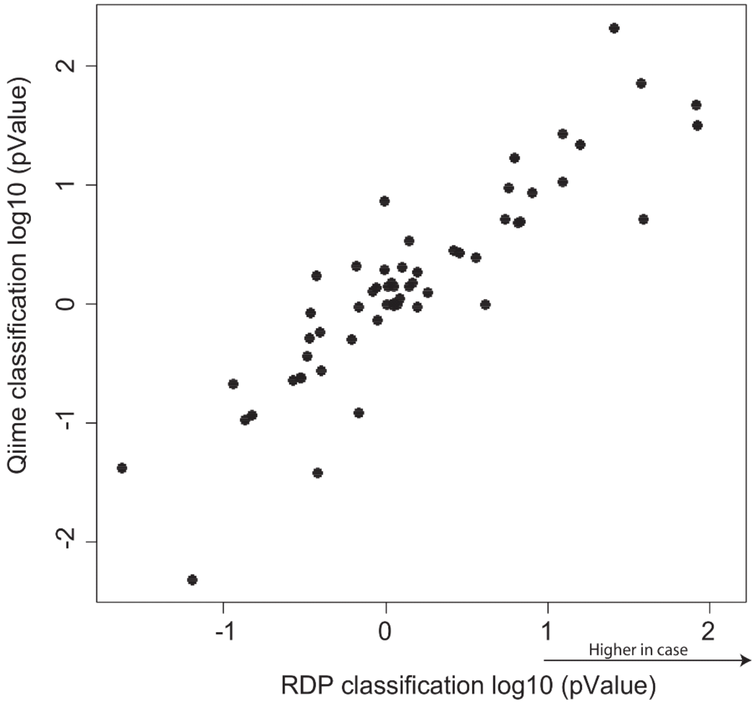

Figure 4.4: Assessment of RDP and QIIME taxonomic classifications using p-values of a t-test between case and control patient status the sign of the p-value was flipped when control is greater than case. General agreement supports the robustness of these findings subject to taxonomic classification method. Figure from Jones *et al*. 2018[104] (Author's own work)

The LEfSe analysis suggested by reviewers did not yield any significant hits when using case versus control status to segment our data. It should be noted that the statistical methodology underpinning LEfSe makes use of the Wilcoxon test, which was already performed as part of our analytic pipeline. The analysis of the microbial co-occurrence networks between case and control status via SPIEC-EASI demonstrated similarity in not only the visual appearance of the case and control microbial networks, but also in graph properties of the networks including degree, natural connectivity and graph similarity, among others (data not shown). Taken together, these observations suggest no strong differences exist between the microbial networks of the case and control samples.

4.6 Discussion

The large size of this study and the small effect sizes of the limited numbers of associations detected suggests that the microbiome is not strongly indicated in the development of diverticula and that there are limited differences in the microbiota of patients with asymptomatic diverticulosis and healthy controls. While convincing evidence of the role of the gut microbiome in the development or severity of gastrointestinal illnesses has been demonstrated numerous times elsewhere, this is not true of previous related findings that were interpreted as being suggestive of a role for the microbiota to play in the development of diverticulosis. For example, a small study of 38 patients discussed visual trends of a decrease in relative abundance of Clostridium IV bacteria for patients with diverticula, as well as shifts in the abundances of Enterobacteriaceae and the ratio of abundances of *Bacteroides* to *Prevotella*, but these results were not statistically significant[105]. This disagreement serves as a reminder of the difficulties in assessing reproducibility between microbiome experiments due to the small effect sizes, sensitivity to sampling and sequencing techniques and differences in potential analytical pipelines. However, our study does differ in that its samples consist of adherent bacteria to biopsies rather than fecal samples, which are likely more tightly associated with the colonic mucosa. As with any negative result, alternative methods may yield differences between the case and control groups, such as via functional assessments of the microbiome.

4.7 Communication of Results

These findings have been published in *Scientific Reports* under the title “An Aberrant Microbiota is not Strongly Associated with Incidental Colonic Diverticulosis” by Jones et al[104].

CHAPTER 5: CASE STUDY OF DAILY CHANGES IN MlCROBIAL COMPOSITION AND DIVERSITY IN THREE PATIENTS WITH ANOREXIA NERVOSA

5.1 Abstract

5.1.1 Background

Previous investigations of anorexia nervosa have revealed limited, but significant microbial associations. Renourishment treatment is frequently prescribed to help return anorexic patients to a healthy BMI, but mechanisms of action and its ultimate effectiveness are still largely unknown. This aim investigates microbial associations with renourishment treatment in a small anorexia cohort.

5.1.2 Materials and Methods

Nutritional measurements of three patients undergoing renourishment treatment for anorexia nervosa were measured alongside fecal samples collected for 16S sequencing. Associations were investigated using models that account for each patient’s characteristic microbial signature.

5.1.3 Results

Significant fluctuations in the abundance of many specific microbes occurred over the course of treatment.

5.1.4 Discussion

While this pilot case study reveals that individual microbial signatures persist through microbial dysbiosis caused by anorexia, larger cohorts are necessary to fully evaluate the associations between microbes and renourishment treatment.

5.2 Introduction and Background

Anorexia nervosa (AN) is a mental illness with high morbidity. In addition to psychiatric and behavioral modifications, it presents many gastrointestinal symptoms. There has been some evidence of association between AN and the dysbiosis of the gut microbiome[106]. Treatment for acute AN typically includes hospitalization and renourishment until the patient reaches a goal BMI. However, the evidence for this method of treatment is not strong and an underlying mechanism is still being sought[107]. While it is generally accepted that the healthy adult microbiome is robust to perturbations over long time periods[8], it is unknown as to whether or not such trends persist in individuals undergoing treatment for AN. Here daily microbiome samples for three patients are assessed alongside energy measurements associated with renourishment treatment in order to evaluate the potential impact of the microbial composition on patient recovery from acute AN.

5.3 Attributions

This work is part of a multi-year study led by Dr. Ian Carroll and Dr. Cynthia Bulik at UNC. My primary role was in the processing of sequencing data and taxonomic classifications, the construction and evaluation of statistical models, and I also contributed to the biological interpretation of the results.

5.4 Materials and Methods

The Biomedical Institutional Review Board at UNC approved this study and all patients, or their guardians provided written consent. For this case study there were three female patients between the ages of 15 and 64 meeting the DSM-5 criteria for the diagnosis of AN. Samples were collected daily by trained medical professionals, and upon release into the partial hospitalization program, patients received training prior to collecting their samples at home with at-home collection kits. Energy intake and weight measurements were conducted daily. Metabolic assessments were made on a weekly basis, and included resting energy expenditure, daily physical activity expenditure, and active energy expenditure.

The V4 region of the 16S rRNA gene was sequenced on an Illumina MiSeq at the High-Throughput Sequencing Facility in the Carolina Center for Genome Sciences at the UNC School of Medicine. The BioProject for this study (PRJNA382889) is available from NCBI. Taxonomic classification utilized QIIME 1.5.0 and RDP version 2.10.1 on the forward reads with a 50% confidence threshold. Post-classification, 141 samples meeting a minimum threshold of 10000 assignments were carried forward for downstream analysis.

The principal model used in the statistical evaluation of this study is an ANOVA linear model which was used to test the both the associations between time undergoing treatment and the patient, including the interaction between these terms. Other models constructed evaluated the associations between microbiome assessments and patient health measures, like BMI, energy intake, etc. Statistical significance in these models was assessed by comparing the full model to a reduced model.

5.5 Results

All patients experienced significant changes in relative abundances at all taxonomic levels over the course of treatment [Figure 5.1, Table 5.1]. This included fluctuations in the relative abundance of consistently detectable taxa as well as taxa which were undetectable for periods of time before returning to detectable levels.

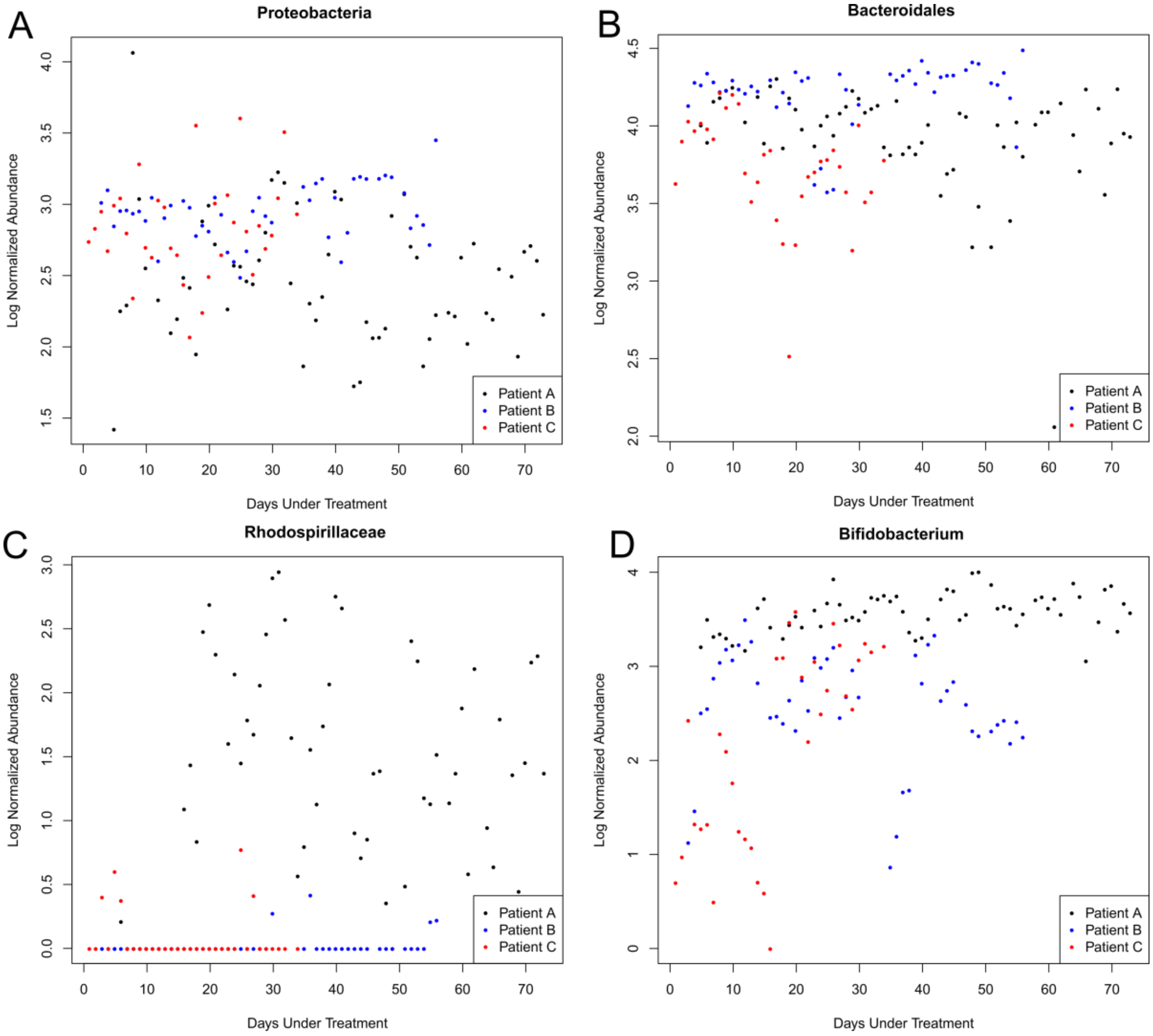

Figure 5.1: Patients preserve individual trajectories during renourishment across taxonomic levels (panels A-D)

Table 5.1: Numerous significant changes to taxa across taxonomic levels Listed from phylum (depicted) to genus over the course of treatment and in a patient-specific manner, as indicated by the significance of the interaction term.

Phylum	**adjANOVA->Day**	**adjANOVA->patient**	**adjANOVA->Day:patient**
Actinobacteria	2.07E-12	1.16E-22	1.59E-02
Bacteroidetes	2.91E-01	1.66E-12	2.32E-02
Crenarchaeota	1.15E-01	3.85E-04	7.80E-01
Cyanobacteria.Chloroplast	9.59E-01	4.56E-03	7.07E-01
Firmicutes	7.61E-03	2.28E-20	2.32E-02
Proteobacteria	1.28E-02	6.73E-09	2.60E-01

Verrucomicrobia	1.50E-04	9.44E-37	2.70E-02

BMI and energy intake were correlated with length of treatment and with each other, so increasing the model complexity by including these terms does not benefit interpretation, especially given the small cohort size [Table 5.2].

Table 5.2: Demonstration of collinearity between length of treatment, BMI and energy intake A similar pattern in significant associations with each individual covariate exists.

Covariate	Phylum	adjANOVA->Covariate	adjANOVA->patient	adjANOVA->Covariate:patient	adjCovariate	adjpatientB	adjpatientC	adjCovariate:patientB	adjCovariate:patientC
Day	Actinobacteria	2.07E-12	1.16E-22	1.59E-02	3.46E-04	1.53E-02	1.37E-03	4.63E-03	9.39E-01
Day	Bacteroidetes	2.91E-01	1.66E-12	2.32E-02	1.44E-02	8.68E-01	2.26E-01	2.93E-02	5.28E-01
Day	Crenarchaeota	1.15E-01	3.85E-04	7.80E-01	6.81E-01	5.55E-01	1.73E-02	6.63E-01	7.07E-01
Day	Cyanobacteria.Chloroplast	9.59E-01	4.56E-03	7.07E-01	7.77E-01	5.55E-01	4.81E-01	4.05E-01	7.07E-01
Day	Firmicutes	7.61E-03	2.28E-20	2.32E-02	6.81E-01	5.55E-01	1.73E-02	2.79E-02	5.28E-01
Day	Proteobacteria	1.28E-02	6.73E-09	2.60E-01	3.93E-01	1.81E-01	4.34E-01	1.60E-01	5.28E-01
Day	Verrucomicrobia	1.50E-04	9.44E-37	2.70E-02	6.81E-01	6.67E-20	1.73E-02	1.46E-02	6.59E-01
BMI	Actinobacteria	2.92E-04	1.22E-13	2.59E-02	2.21E-02	4.19E-02	8.08E-01	2.24E-02	7.47E-01
BMI	Bacteroidetes	3.01E-07	3.42E-06	2.59E-02	2.36E-01	2.17E-01	6.22E-01	1.43E-01	2.36E-01
BMI	Crenarchaeota	3.26E-03	9.96E-03	7.87E-01	2.80E-01	4.65E-01	6.31E-01	4.81E-01	7.02E-01
BMI	Cyanobacteria.Chloroplast	9.87E-01	2.22E-02	7.87E-01	6.23E-01	6.79E-01	8.08E-01	5.83E-01	7.47E-01

BMI	Firmicutes	9.10E-16	3.82E-06	2.59E-02	7.05E-01	4.19E-02	6.31E-01	2.24E-02	6.82E-01
BMI	Proteobacteria	9.74E-01	8.43E-05	7.87E-01	8.68E-01	7.50E-01	8.08E-01	5.83E-01	7.02E-01
BMI	Verrucomicrobia	6.37E-06	7.42E-28	1.23E-01	7.05E-01	1.15E-02	6.31E-01	5.41E-02	7.02E-01
Energy Intake	Actinobacteria	1.50E-10	5.22E-24	5.47E-03	1.05E-04	5.04E-01	8.64E-01	1.15E-03	3.70E-01
Energy Intake	Bacteroidetes	6.29E-03	8.12E-12	3.27E-02	1.65E-02	4.66E-01	6.71E-01	2.73E-02	8.20E-01
Energy Intake	Crenarchaeota	8.76E-01	8.10E-05	7.01E-01	5.09E-01	5.84E-01	1.65E-01	5.95E-01	5.70E-01
Energy Intake	Cyanobacteria.Chloroplast	8.76E-01	3.90E-03	7.75E-01	5.09E-01	8.62E-01	6.71E-01	5.95E-01	8.50E-01
Energy Intake	Firmicutes	8.76E-01	3.26E-21	3.27E-02	5.09E-01	5.84E-01	6.71E-01	5.44E-02	5.70E-01
Energy Intake	Proteobacteria	1.18E-01	2.00E-09	7.01E-01	5.09E-01	8.57E-01	8.64E-01	4.03E-01	5.70E-01
Energy Intake	Verrucomicrobia	4.91E-07	9.67E-36	5.77E-02	5.09E-01	3.87E-08	1.65E-01	2.73E-02	5.70E-01

Weekly metabolic measures were similarly highly correlated with length of treatment and did not yield significant associations with microbial measures (data not shown).

5.6 Discussion

While the small cohort size cautions against the over-interpretation of these findings, the data suggests that even in the case of acute AN, the microbial signatures of individual patient’s persist. This result suggests that similar longitudinal studies be conducted in other gastrointestinal illnesses to investigate whether or not the disease signature "overwhelms" the individual microbial signature. These findings are in agreement with other longitudinal studies of gut microbiomes that showed periods of stability and volatility, but similarly they did not reveal evidence for a common core of microbiota shared between individuals[8]. Also in agreement with other studies, this cohort did not uncover strong associations between BMI and the gut microbiome[108]. These limited results still remain promising and suggest that larger cohorts with more frequent collection of metabolic assessment data, as well as functional assessments of the microbiome through whole-genome sequencing, may be able to explore microbial mechanisms underlying weight dysregulation associated with AN and its treatment via renourishment.

5.7 Communication of Results

These findings have been published in *European Eating Disorders Review* under the title "Daily Changes in Composition and Diversity of the Intestinal Microbiota in Patients with Anorexia Nervosa: A Series of Three Cases" by Kleiman et al[23].

CHAPTER 6: INTERACTIONS OF SEX AND STRESS STATUS MODULATE THE MICROBIOME IN A MOUSE MODEL

6.1 Abstract

6.1.1 Background

Human males and females differ in the frequency and presentation of mental illnesses such as anxiety and depression. Unfortunately, experimental cohorts in animal models of human disease often fail to include female animals. Here the microbiota-gut-brain axis is evaluated in a mixed sex mouse cohort.

6.1.2 Materials and Methods

32 mice (16 male, 16 female) were split between a control group and a stress experimental group that was subject to alternating days of physical stress during the study period. Fecal and cecal samples were collected for 16S rRNA sequencing, and assessments from three behavioral tests measuring anxiety were evaluated against individual taxa and microbial composition shifts.

6.1.3 Results

Mixed linear models revealed significant shifts in microbial community with respect to sex, stress-status, and the interaction term of these fixed effects. Similarly, individual taxa were differentially abundant with respect to these model terms. No significant associations were seen between individual behavioral measures and specific microbes.

6.1.4 Discussion.

Many of the taxa seen to be differentially abundant are in agreement with previous studies that have evaluated sex differences or stress differences within the same sex in mouse. The presence of a significant interaction term can be interpreted as an indicator

that the sex of the mouse somehow modulates the response of the animal to stress leading to different changes of specific microbes and the community composition as a whole. These results suggest that future studies of the microbiota-gut-brain axis use mixed sex cohorts.

6.2 Introduction and Background

Experiments using animal models, even those used in pharmaceutical pipelines assessing drug efficacy and safety, frequently fail to include female animals[109,110]. In 2014 the NIH Director and the Director of the Office of Research on Women's Health began to develop a strategy to balance male and female model animals in funded research[111]. However, it is still currently the case that grants lacking sex as a biological variable continue to receive acceptable scores by reviewers[112]. This cohort bias is particularly troublesome in animal studies modeling human mental illnesses[113], as conditions such as anxiety and depression are known to differ in frequency between men and women, as well as in presentation[114,115]. There are also known differences in differences in male/female drug response in humans[116–118]. Such results are perhaps unsurprising given what is known about differences in the male/female psychophysiological responses to stress. In a similar manner, animal models have also been shown to exhibit sex-dependent responses to pharmaceuticals[119,120]. However, the negative consequences of ignoring sex differences can also extend to drugs where sex-based differences in results are naively thought to be unexpected, such as antibiotics[121]. Indeed, understanding potential sex-dependent differences with regards to pharmaceutical use may become all the more important to properly conducting metagenomics studies given recent results that indicate that many non-antibiotic pharmaceuticals significantly influence the microbiome[122].

Chronic stressors have been shown by others to impact microbial diversity and differential abundances[123]. Potential sex-related differences in microbial abundances in response to stress and anxiety also underline the bi-directionality of the microbiota-gut-brain axis. The ability of docosahexaenoic acid (DHA) to alleviate anxiety or depression-like symptoms for male mice and not female mice would have remained unnoticed if it were not for the mixed sex cohort being used[124]. Similarly, a study of a mouse model of autism using BTBR mice recently enjoyed success by creating autism-trait microbial profiles because of employing a mixed-sex cohort[125].

This aim investigates the microbiota-gut-brain axis through assessing shifts in the microbiome associated the interaction of the sex of the host with the stress response. Associations between the changes in the microbiome and the changes in behavior due to stress protocols are also evaluated.

6.3 Attributions

This research was done in collaboration with Dr. Mark Lyte of Iowa State University. My primary roles in this project were in the processing of taxonomic classifications, the development and evaluation of statistical models, and the biological interpretation of results.

6.4 Materials and Methods

A cohort of 32 six-week-old CF-1 mice was purchased from Charles River Laboratories. There were 16 male and 16 female animals, and these were same-sex-housed 4 to a cage. This led to 8 total cages with 2 cages for the stress group and 2 cages for the control group per sex. Stress was induced by alternating days of physical restraint and a forced swimming challenge. At the end of the 19-day testing period, the mice underwent behavioral evaluations in the form of three behavioral tests to evaluate for anxiolytic behavior. These tests were the light-dark box, the elevated-plus maze and the open field. This resulted in 39 behavioral characterizations in total across the three tests. At the end of behavioral testing the animals were sacrificed, and various tissues were collected including fecal and cecal content samples for 16S rRNA gene sequencing and blood for measures of hormones such as the stress-associated hormone corticosterone. These experiments were performed at Texas Tech University by the same female lab technician[126]. All experiments approved by the Institutional Animal Care and Use Committee of Texas Tech University Health Sciences Center.

DNA sequences were isolated using the PowerSoil DNA Isolation Kit and quantified using a Qubit 2.0 Fluorometer. The sequences were amplified using the primer set from the Earth Microbiome Project for the V4 region of the 16S rRNA gene (515F-806R) and then quantified using the Qubit. The sequencing was completed using the Illumina MiSeq platform by facilities at the Argonne National Laboratory.

The behavioral and sequencing data were evaluated using nonparametric and parametric statistical models. A series of mixed linear models were created to assess various fixed effects of explanatory variables such as the sex of the mouse, experimental group (stress/control), and behavioral scores on the relative abundances of microbes, with the random effect being the cage grouping of the animals. Response variables included the scores of individual behaviors, log normalized abundance of specific OTUs, the alpha-diversity of the microbial community as measured by Shannon diversity, and the principal coordinate axes from the beta-diversity as determined by the Bray-Curtis dissimilarity.

6.5 Results

There are several significant associations between the terms of the sex*stress model and individual measures of anxiolytic behavior [Figure 6.1].

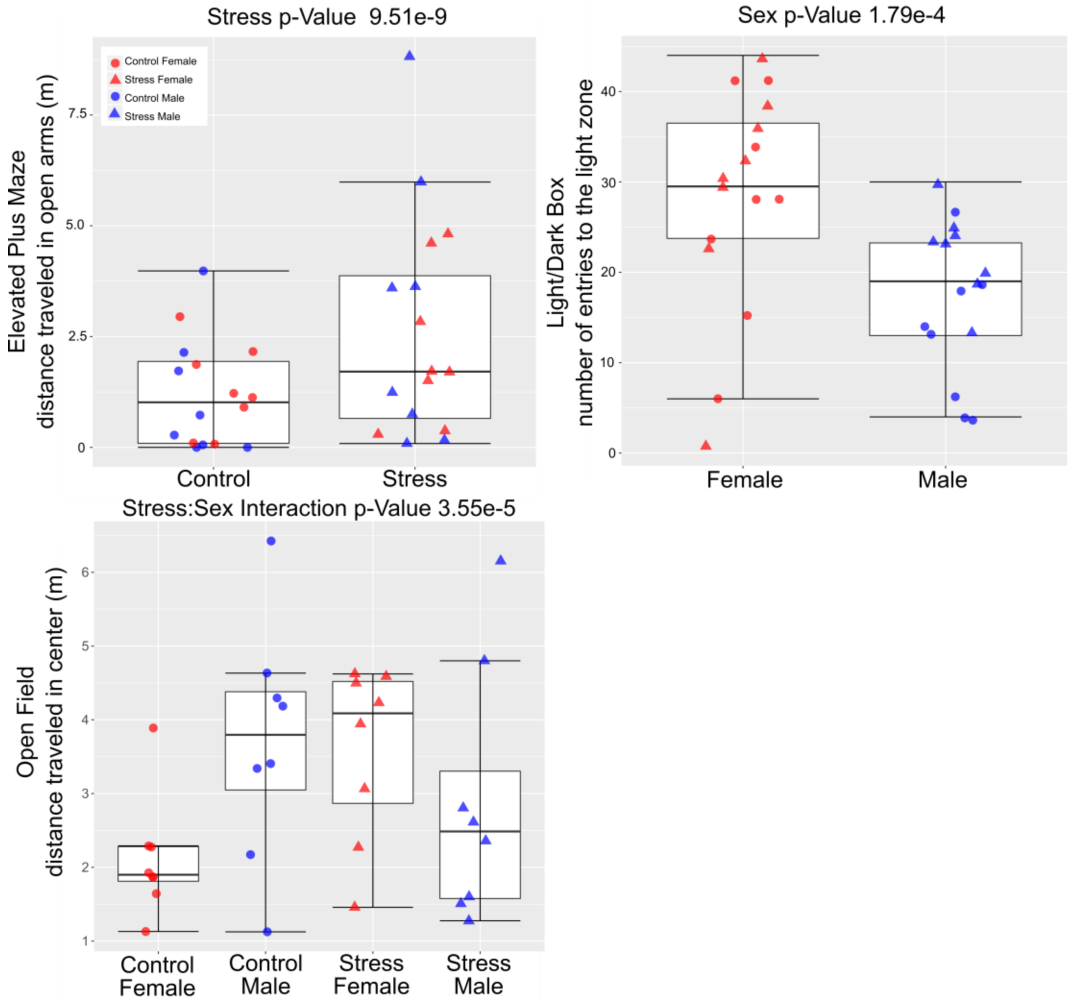

Figure 6.1: An example subset of significant associations between specific behavioral measures and model terms of stress, sex and sex*stress from a mixed linear model. These results can be interpreted as the induction of a significant shift in behavior due to experimental stress protocols.

However, the measurements of the stress hormone corticosterone were not significantly different between stress and control animals [Figure 6.2].

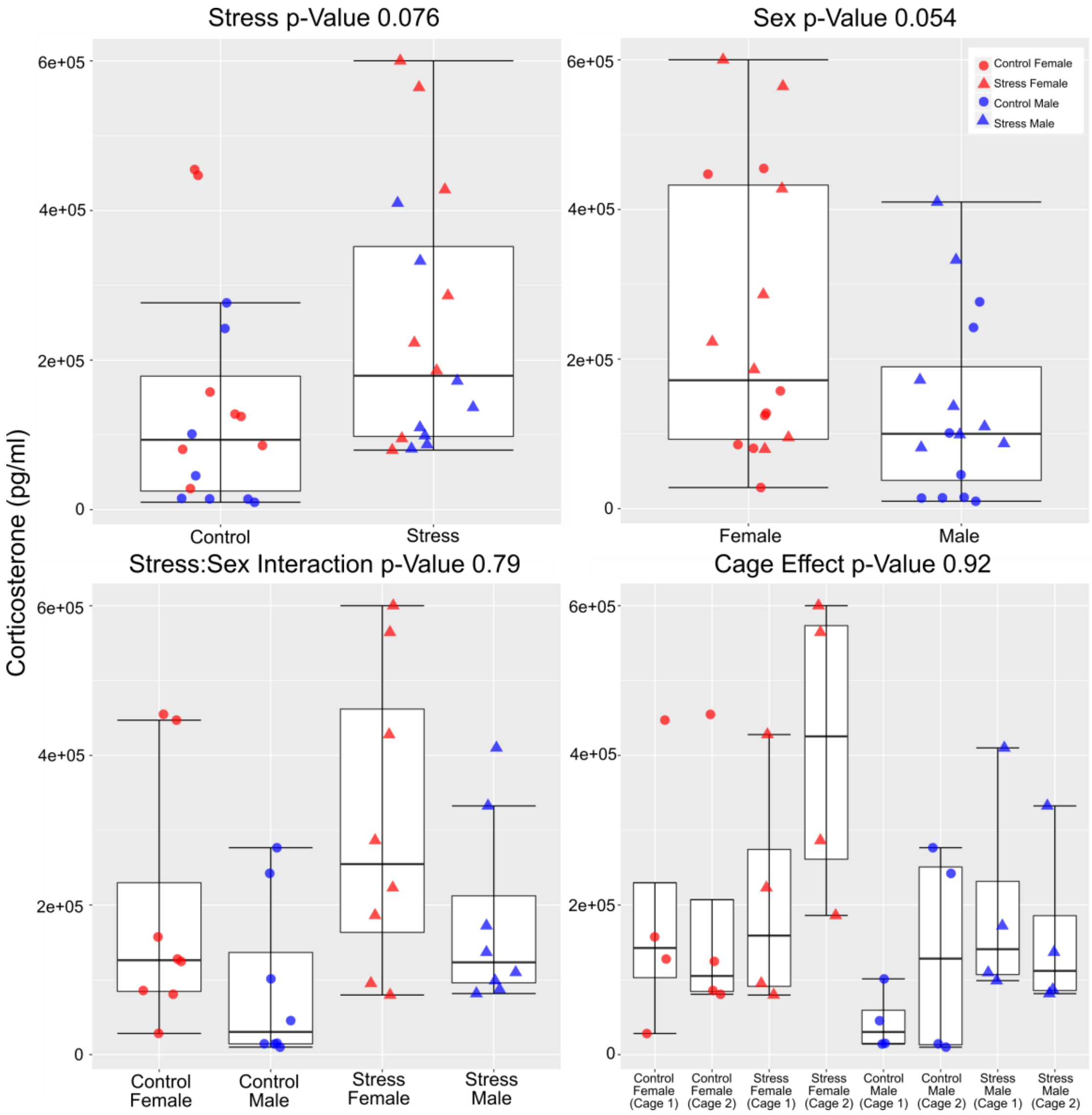

Figure 6.2: Using similar mixed linear models with concentrations of the stress hormone corticosterone as the response variable yielded trends for the stressed animals and the female animals having higher concentrations, but these were not significant at the 0.05 level.

There were many significant differences when the microbiome data was viewed at the community level and when assessing specific OTUs. The ordination yielded MDS axes that separated the animals on each of the terms of the mixed linear model [Figure 6.3].

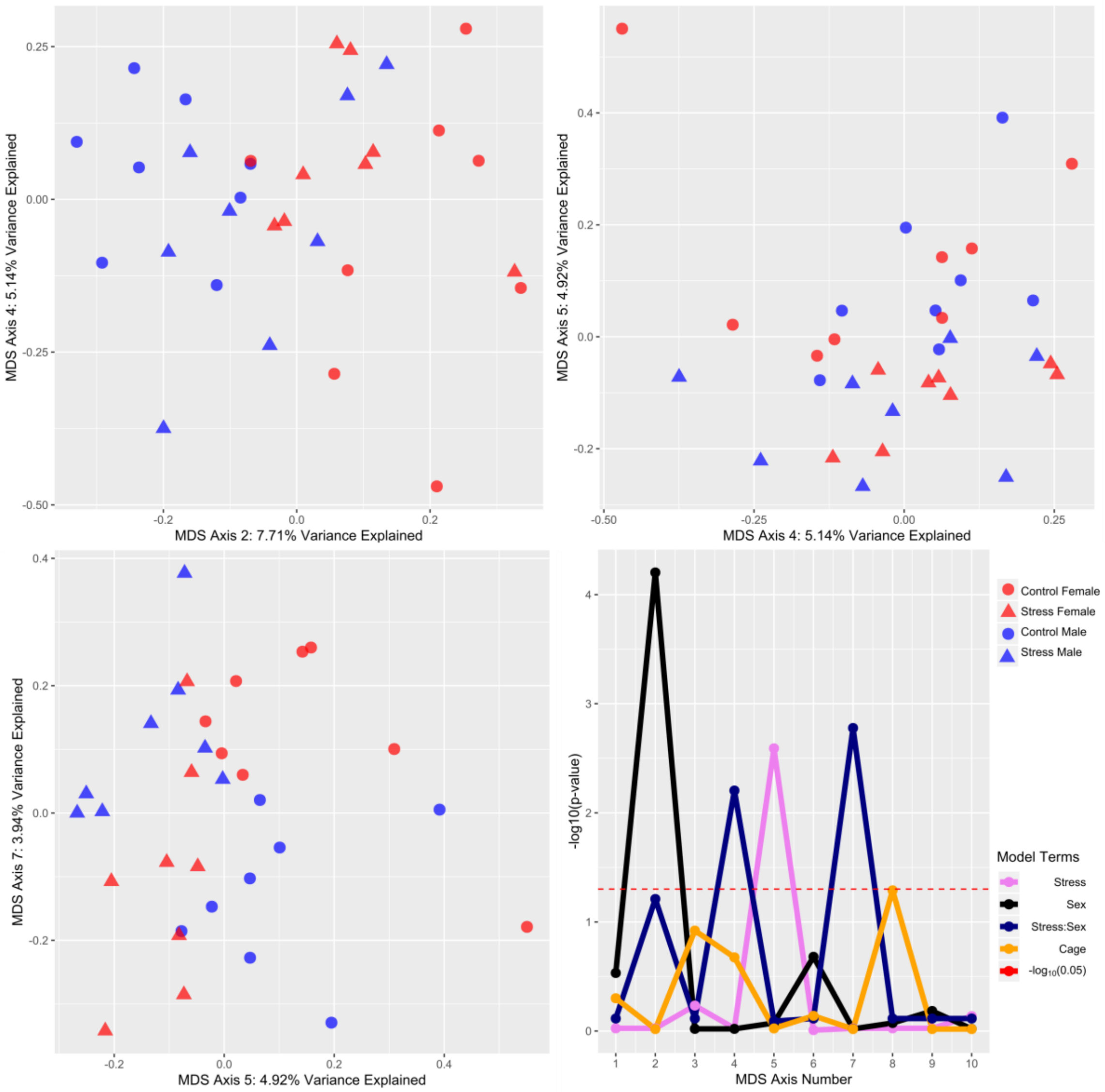

Figure 6.3: MDS-ordination axes are significantly associated with the sex, stress, and sex:stress model terms of the mixed linear model. Top left, top right, and bottom left: ordination plots of axes associated with sex (axis 2), sex:stress (axes 4 and 7) and stress (axis 5). Bottom right: Plot of -log10(p-value) showing the significant axes from the ordination. Cage is never simultaneously significant with the model terms of interest, indicating that cage effect does not drive these significant differences. The dashed line represents the transformation of the significance threshold of 0.05.

The male mice had higher Shannon diversity values than the females, but this trend was not seen with respect to the other terms in the model. Numerous OTUs were also significantly differentially abundant for the sex, stress and sex:stress terms of our mixed linear model [Table 6.1].

Table 6.1: Table of significant genera and OTU identified labeled by term in the mixed linear model

Genus (OTU)	**Wilcox -Sex**	**Wilcox - Stress**	**Sex**	**Stress**	**Sex:Stress Interaction**	**Cage**
Adlercreutzia (631764)	0.02	0.57	< .001	< .001	< .001	0.49
Odoribacter (170335)	0.02	0.57	< .001	0.001	0.003	0.94
Bacteroides (197537)	0.71	0.98	0.3	0.95	0.002	0.59
Bacteroides (198449)	0.16	0.52	0.002	0.008	0.012	0.84
AF12 (190026)	0.07	0.82	< .001	0.31	0.45	0.64
Lactobacillus (539647)	0.75	0.94	0.42	0.95	0.02	0.75
Lactobacillus (47365)	0.86	0.44	0.6	0.006	0.11	0.86
Lactobacillus (343431)	0.87	0.59	0.99	0.003	0.93	0.59
Clostridium perfringens (323526)	0.19	0.57	0.014	0.13	0.7	0.94
Sarcina (446153)	0.85	0.85	0.99	0.75	0.02	0.98
Ruminococcus gnavus (family Lachnospiraceae) (269107)	0.07	0.99	< .001	0.32	< .001	0.11
Ruminococcus gnavus (family Lachnospiraceae) (348336)	0.90	0.77	0.87	0.62	0.008	0.77
Ruminococcus gnavus (family Lachnospiraceae) (352008)	0.22	0.95	< .001	0.9	0.012	0.49
Anaerostipes (534926)	0.07	0.79	< .001	0.18	0.39	0.66
Coprococcus (4476330)	0.73	0.98	0.065	0.81	0.001	0.31
Coprococcus (269828)	0.82	0.85	0.74	0.82	0.035	0.94

Coprococcus (1107439)	0.08	0.85	0.015	0.71	0.74	0.96
Oscillospira (276386)	0.08	0.96	< .001	0.75	0.92	0.63
Oscillospira (387615)	0.71	0.12	0.99	0.014	0.95	0.97
Ruminococcus (Ruminococcaceae) (320224)	0.90	0.96	0.83	0.9	0.008	0.59
Ruminococcus (family Ruminococcaceae) (339031)	0.88	0.76	0.83	0.99	0.02	0.87
Ruminococcus (family Ruminococcaceae) (405780)	0.02	0.95	0.029	1	0.95	0.94
Coprobacillus (4449984)	0.06	0.93	0.019	0.79	0.19	0.97
Anaeroplasma (835872)	0.02	0.52	< .001	0.099	0.29	0.98

However, there were no significant associations between individual behavioral measures and specific OTUs. The results from the microbial communities from the fecal and cecal samples are in broad agreement [Figure 6.4].

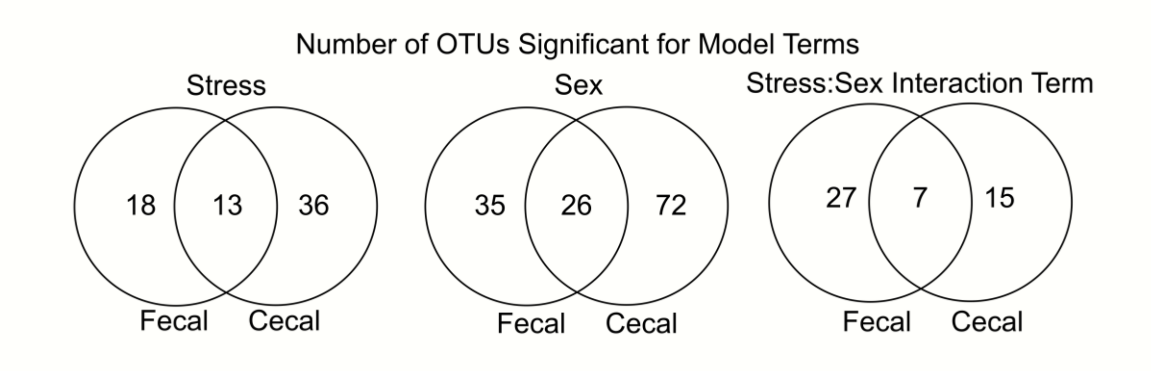
Number of OTUs Significant for Model Terms
Stress
18
13
36
Fecal
Cecal
Sex
35
26
72
Fecal
Cecal
Stress:Sex Interaction Term
27
7
15
Fecal
Cecal

Figure 6.4: Extent of overlap in significant genera between fecal and cecal samples collected and sequenced.

6.6 Discussion

The significant associations between individual behaviors and terms in the model can be taken as indicative of a successful stress protocol in terms of inducing a stress response. The lack of significant association between corticosterone and the experimental protocol may help to explain the lack of significant associations between specific microbes and specific behavioral measures indicating a sub-threshold activation of the HPA-axis. Several of the taxa observed to be significantly differentially abundant in this study have also been detected in other studies involving sex-based differences or stress-based differences. For example, members of Lachnospiraceae are elevated in the stress cohort, and this response to stress has been seen in an all-female cohort subjected to a similar stress protocol[127]. The relative abundance patterns of Clostridium perfringens, a known pathogen, agree with previous work that has shown it to be elevated in abundance in both female animals and human females[128]. Importantly, this experiment enabled the evaluation of the sex:stress interaction term and has indicated several taxa which may be involved in modulating the stress response due to sex. However, it should be cautioned that studies such as this one can only detect associations and not causality. While there were no significant associations between individual behavioral measurements within the three tests and specific microbes, this is not surprising given the large number of tests for which to control for the false discovery rate and the small effect sizes of microbial associations with human health covariates seen in large human cohorts[15,16]. One of the broad-reaching conclusions of this work is that the presence of a significant sex:stress interaction term in the statistical models of the behavioral and microbiome data suggests that future studies of the gut-brain-microbiota axis should always be conducted using mixed sex cohorts.

A follow-up study, while outside of the scope of this dissertation, includes a replication cohort and other mixed sex mouse cohorts with additional forms of stressors, like sleep deprivation and social stress, and their corresponding behavioral measurements. These are complemented by a more extensive assay of hormone and neurochemical concentrations to enable a more function-oriented perspective to changes within the microbiota-gut-brain axis under various stressors. The initial results of analyzing this data suggest the replicability of the microbiome results observed in this aim across cohorts.

6.7 Communication of Results

Elements of this work were presented at the UNCC Graduate Research Symposium in 2017, and as a poster at ASM Microbe 2018 in Atlanta, Georgia. This work has been published as “Interactions Between Stress and Sex in Microbial Responses Within the Microbiota-Gut-Brain Axis in a Mouse Model” in *Psychosomatic Medicine* 2018 May;80(4):361-369 by Tsilimigras et al[129].

CHAPTER 7: MICROBIAL COMMUNITY COMPOSITION, ANTIBIOTIC CONCENTRATIONS AND ANTIBIOTIC RESISTANCE GENES UPSTREAM, DOWNSTREAM AND WITHIN A NORTH CAROLINA URBAN WATER SYSTEM

7.1 Abstract

7.1.1 Background

Wastewater treatment plants are primarily evaluated on their removal of known pathogens. Here their ability to remove antibiotic resistance genes is evaluated.

7.1.2 Materials and Methods

Whole genome sequencing was used to classify the organisms present and evaluate their antibiotic resistance genes against a targeted database. These relative abundances were compared pairwise across sampling locations.

7.1.3 Results

The microbial community downstream of the wastewater treatment processing was largely restored to its upstream composition. Specific pathogens and most antibiotic resistance genes are not significantly elevated in samples downstream of treatment. Concentrations of several antibiotics remain significantly elevated after processing.

7.1.4 Discussion

The wastewater treatment plants performed their specified task, but the elevated concentrations of antibiotics post-treatment suggest that further studies could be developed to establish treatment protocols that better support antibiotic stewardship.

7.2 Introduction and Background

Wastewater treatment plants have historically been assessed by their ability to remove pathogenic organisms[130]. However, their contributions to the accumulation and concentration of pharmaceuticals in the ecosystem is becoming increasingly appreciated[131,132]. This also means that wastewater treatment practices have an underappreciated role in antibiotic stewardship. This has the potential to limit the future utility of pharmaceuticals like antibiotics as the wastewater treatment process could lead to the evolution and dissemination of antibiotic resistance genes through horizontal gene transfer[133]. This is particularly troubling because there are numerous signs that we may be in the beginnings of an antibiotic resistance crisis where the ability of microbes to evolve and spread resistance genes is outpacing the development of novel antibiotics[134]. This situations is due, in large part, to the market failure on the part of pharmaceutical companies to develop antibiotics without strong guidance and incentives from governmental and nonprofit organizations[135,136]. The accumulation of pharmaceuticals like antibiotics can also potentially lead to problems in the wastewater treatment process itself if the microbes added to properly degrade sewage are sensitive to such compounds[122,137].

Recently a large study was conducted by UNC Charlotte faculty in collaboration with the sequencing resources at the David H. Murdock Research Institute (DHMRI) to collect many water samples upstream, downstream and within Charlotte’s Mallard and Sugar Creek wastewater treatment facilities. These samples were used to generate whole-genome metagenomic sequencing data and this aim evaluates the changes to microbial communities before, during and after wastewater treatment, as well as to investigate the presence of antibiotic resistance genes throughout the process. In addition, the concentrations of several common antibiotics were measured to investigate the impact wastewater treatment had on downstream levels of common pharmaceuticals, and potential interactions between microbes and antibiotic concentrations.

7.3 Attributions

This work represents part of a larger study led by Dr. Cynthia Gibas. My role in this project was to develop the analysis pipelines for the statistical inferences related to the microbial communities, taxa-by-taxa differential abundance, antibiotic resistance gene abundances, and antibiotic concentrations. I oversaw and was assisted by James Johnson and Dr. Anju Lulla in these analyses which rely on the results of bioinformatics pipelines constructed by Dr. Kevin Lambirth, Dr. Lulla, and Abrar Al-Shaer. I also assisted Dr. Lambirth, Dr. Gibas and others in the biological interpretation of these results. Dr. Lambirth collected and processed all samples prior to sequencing at DHMRI.

7.4 Materials and Methods

7.4.1 Experimental Design

Water samples were collected with three technical replicates each at four timepoints at locations near hospital and residential areas, upstream and downstream of the wastewater treatment plants, and at multiple locations within both the Mallard Creek and Sugar Creek wastewater treatment facilities. For Mallard Creek, these sites were raw influent (INF), primary clarifier influent (PCI), primary clarifier effluent (PCE), aeration tank effluent (ATE), and final clarifier effluent (FCE). For the Sugar Creek plant, the PCI point was not able to be sampled, but the ultraviolet disinfected effluent (UV) was able to be sampled. Additionally, samples were taken in two rural streams far from Charlotte (one in the Appalachian mountains and one in Uwharrie forest) as controls for the fourth and final timepoint in mid-summer. This yielded a total of 66 samples for the first three timepoints and 78 samples for the last timepoint. These locations are depicted in [Figure 7.1].

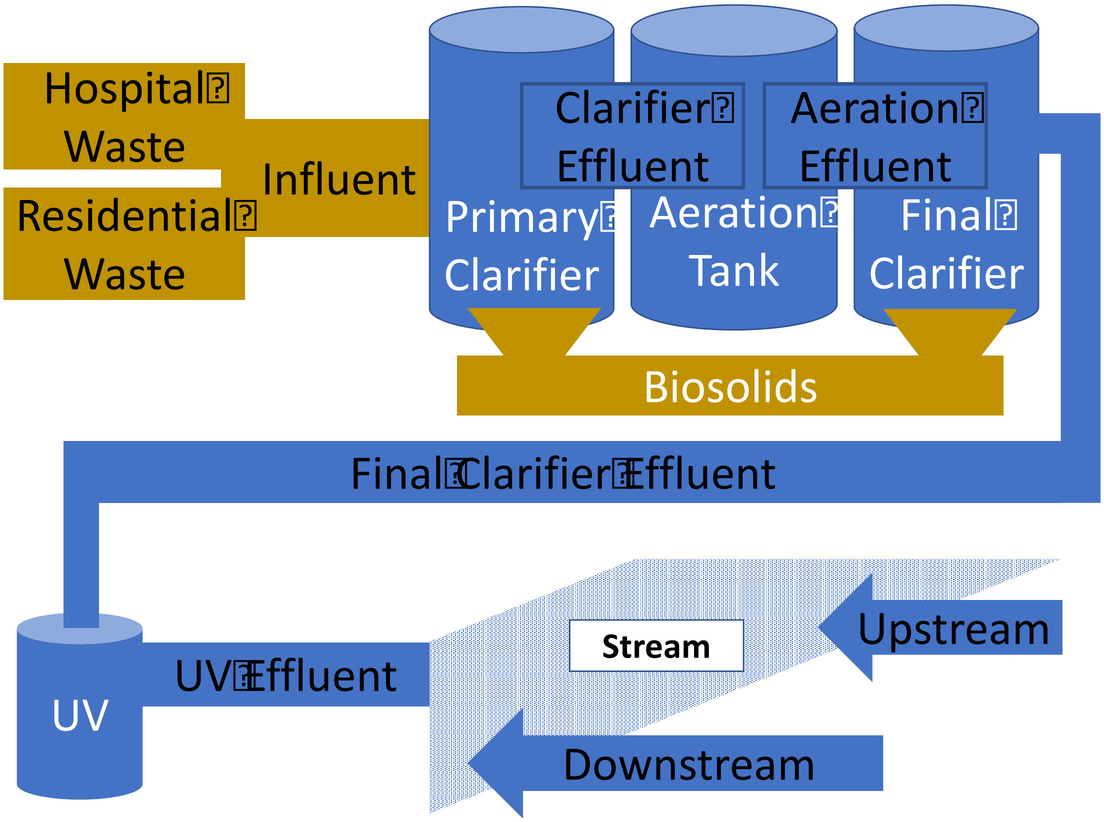

Figure 7.1: Diagram of relative location of sampling sites within and before and after the wastewater treatment plants. Figure from Lambirth *et al.*[138] (Author's own work).

7.4.2 Sequencing and Sequence Pre-Processing

These samples were then sequenced at DHMRI using both 16S rRNA and WGS technologies on Illumina HiSeq 2500 lanes to investigate the microbial communities. WGS sequences were trimmed for quality assessments using Trimmomatic[139] before forward and reverse reads were merged using PEAR[140]. The technical replicate that yielded the highest sequencing depth was used as the representative sample in the statistical analyses.

7.4.3 Taxonomic Classification and Antibiotic Resistance Gene Profiling

Taxonomic classifications for WGS sequencing data came from merged reads classified using MetaPhlAn2[52] using the default settings. The determination of antibiotic resistance markers came from the ShortBRED[141] pipeline using a custom database constructed from the Comprehensive Antibiotic Resistance Database (CARD)[142] and the Lahey Clinic beta-lactamase database.

7.4.4 Statistical Modeling

Methods of statistical inference include linear regression models where the response variable for the taxonomic classifications were either the microbial abundance of specific taxa, and alpha-diversity or beta-diversity measures of community composition calculated using the R package vegan[66]. Other response variables came from the antibiotic resistance gene abundances from ShortBRED and the antibiotic concentrations. Explanatory variables in the model included terms of location (Mallard or Sugar Creek), sample site (where in the waterway the sample came from), and the timepoint at which the sample was collected. These models were used to compare sites in a pairwise manner. This means that the general format of the statistical models used two categories for Sugar or Mallard Creek, two categories for the sample site, and four categories for the timepoint. These model terms were taken as additive only since models with interaction terms did not achieve many significant differences. The small number of timepoints and the irregular intervals between collection dates led to the treatment of the different timepoints as categorical differences rather than terms requiring sophisticated linear models with temporal autocorrelation, which could not likely be justified by the small size of subsets of the dataset being compared. Additionally, associations between antibiotic concentrations and the relative abundance of specific microbes were investigated using standard correlation tests.

7.5 Results

7.5.1 Microbial Communities Seen Via Beta-Diversity

The investigations of the beta-diversity via Bray-Curtis PCoA showed a clear separation in the microbial communities of sampling sites involved in the wastewater treatment process as compared to stream sites pre- and post-treatment [Figure 7.2].

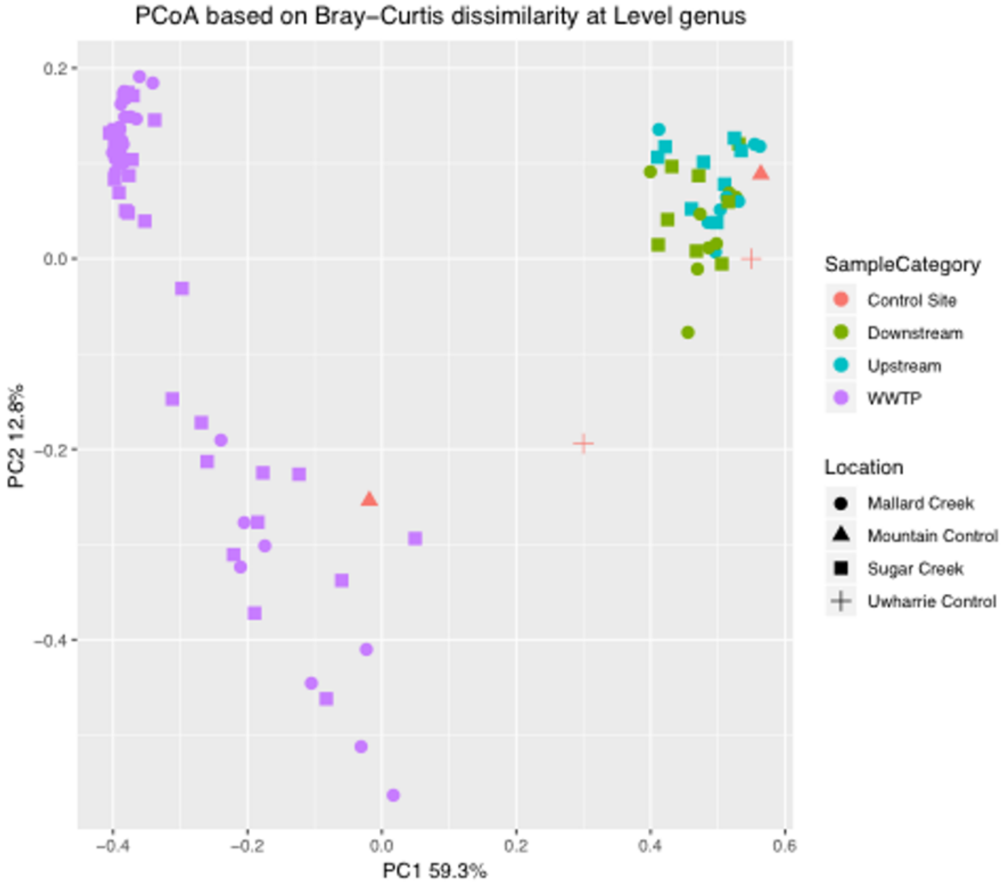

Figure 7.2: Ordination of different sample subsets Results visually indicate that stream communities downstream resemble upstream communities after treatment. Figure from Lambirth *et al*.[138] (Author's own work).

7.5.2 Differences in Antibiotic Concentrations During Processing Steps

The concentrations of several antibiotics were elevated at the sampling sites downstream of the wastewater treatment plant [Figure 7.3].

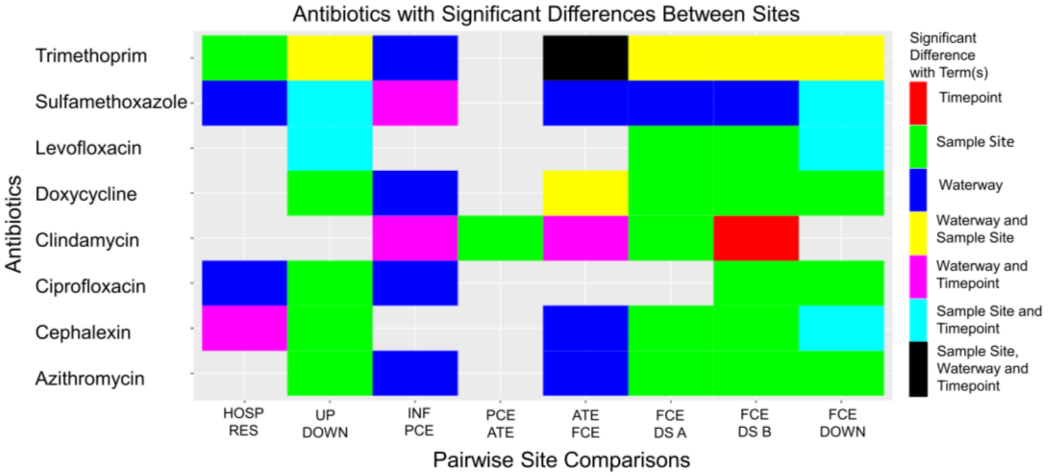

Figure 7.3: Several antibiotic concentrations were significantly elevated between the upstream and downstream sampling sites. Figure from Lambirth *et al*.[138] (Author’s own work).

7.5.3 Bacterial Differences in Abundance During Treatment Stages

In contrast to the community and antibiotic concentration differences, most potential pathogens were not significantly different between upstream and downstream sites, though there were significant differences in taxonomic abundances throughout points of comparison within the treatment process [Table 7.1].

Table 7.1: Significant differences in taxa for the site-site comparisons. Note that few taxa have significant differences between the upstream and downstream samples, especially in comparison to the other site-pairs investigated. Table from Lambirth *et al*.[138] (Author’s own work).

Taxon	***p* Value**	**Higher Abundance**
Peptostreptococcaceae	0.0039	Downstream to Upstream
Afipia	0.0093	Downstream to Upstream
Holospora	0.0039	Downstream to Upstream
Azoarcus	0.0114	Downstream to Upstream
Acinetobacter	0.013	Downstream to Upstream

Bppunalikevirus	0.0093	Downstream to Upstream
Yualikevirus	0.0096	Downstream to Upstream
Sphingobium	0.01	Rural to Upstream
Kocuria rhizophila	0.0318	Residential to Hospital
Nitrospira defluvii	0.0216	ATE to PCI
Caulobacter sp	0.0058	ATE to PCI
Afipia clevelandensis	0.0048	ATE to PCI
Rhodopseudomonas paulustris	0.012	ATE to PCI
Hyphomicrobium denitrificans	0.0114	ATE to PCI
Mesorhizobium sp	0.0183	ATE to PCI
Paracoccus sp	0.0439	ATE to PCI
Reyranella massiliensis	0.0111	ATE to PCI
Sphingobium xenophagum	0.0184	ATE to PCI
Sphingopyxis sp	0.0003	ATE to PCI
Alicycliphilus sp	0.0004	ATE to PCI
Limnohabitans sp	0.0005	ATE to PCI
Polaromonas sp	0.0003	ATE to PCI
Variovorax sp	0.0014	ATE to PCI
Azoarcus sp	0.0006	ATE to PCI
Dechloromonas sp	0.011	ATE to PCI
Methyloversatilis sp	0.0008	ATE to PCI
Thauera aminoaromatica	0.0212	ATE to PCI
Actinobacter parvas	0.025	ATE to PCI
Turneriella parva	0.0058	ATE to PCI
Methanobrevibacter sp	0.0357	ATE to PCI
Gordonia amarae	0.0476	ATE to PCI
Tetrasphera elongata	0.0218	ATE to PCI
Rhodococcus	0.0409	Downstream to FCE
Actinobacterium sp	0.0116	Downstream to FCE
Polynucleobacter necessarius	0.00000007	Downstream to FCE
Limnohabitans	0.00000007	Downstream to FCE
Methylotenera	0.0404	Downstream to FCE
Bppunalikevirus	0.0132	Downstream to FCE
Yualikevirus	0.0266	Downstream to FCE

7.5.4 Differences in Antibiotic Resistance Genes Between Sites

The small number of significant differences between upstream and downstream sites indicates that the treatment process does not introduce many new antibiotic resistance genes to the bacterial populations upon release of the treated water.

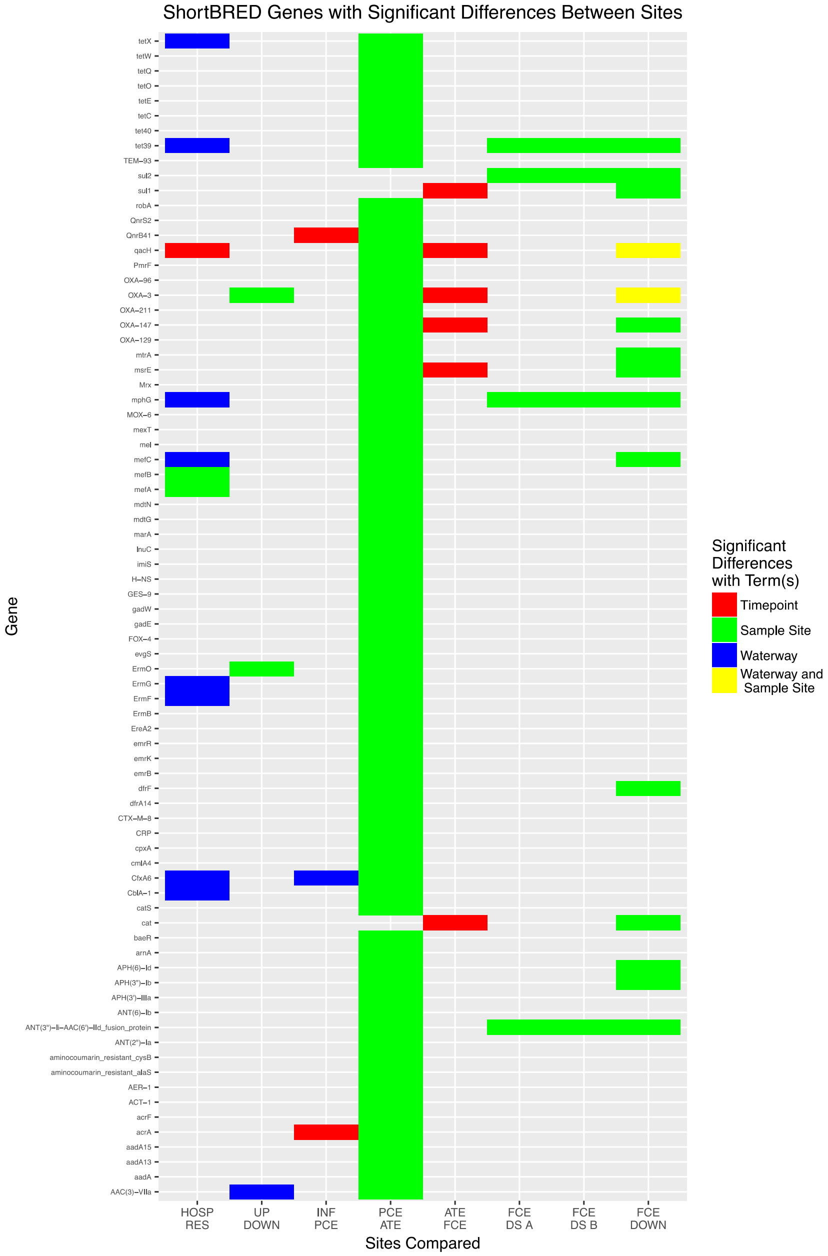

Figure 7.4: Antibiotic resistance genes as predicted by ShortBRED colored by the model terms significantly different between pairs of sampling sites. Figure from Lambirth *et al.*[138] (Author’s own work).

7.5.5 Associations Between Specific Taxa and Antibiotic Concentrations

There was no evidence of strong associations with significant p-values between antibiotic concentrations and specific taxa (data not shown).

7.6 Discussion

The lack of significant increases in microbial relative abundances between upstream and downstream sites can be interpreted as the treatment plants are fulfilling their established goals of removing pathogens in the course of the treatment process prior to release. The relative abundance of antibiotic resistance elements was also not broadly elevated downstream of processing, though wastewater treatment plants are not routinely assessed in their ability to remove antibiotic resistance genes themselves, other systems have failed to do so[143,144]. However, the exposure to antibiotic concentrations, even at sub-lethal levels, has been implicated in driving the evolution and dissemination of antibiotic resistance elements[145,146]. This possible means of driving antibiotic resistance remains even though we did not observe any strong associations between microbial relative abundances and concentrations of antibiotics measured as the water samples taken only capture transient interactions between antibiotic and microbe. At the level of the microbial community, the samples collected downstream of the treatment process largely resembles the upstream communities, and the communities more closely resemble stream samples as the wastewater processing proceeds.

Future work could include a careful assessment of the attenuation of human gut and wastewater treatment associated microbes downstream of processing via systematic sample collection and quantitative microbial load measurements (i.e. qPCR), which may help to guide human usage and interactions with proximal downstream waterways. Other recent studies have revealed that even non-antibiotic pharmaceuticals can significantly impact microbial communities[92], and this suggests that the expansion of small molecule detection to include such compounds may yield insight into the microbial interactions with these molecules in environmental settings after the initial interactions within the host. Ongoing work within this project includes similar investigations of antibiotic resistance in other types environmental samples like soil and sediment.

7.7 Communication of Results

This work was presented at the 2018 North Carolina Microbiome Consortium Symposium at Research Triangle Park, NC on May 15th, 2018. It has been published in the journal *Water* under the title "Microbial Community Composition and Antibiotic Resistance Genes Within a North Carolina Urban Water System" by Lambirth et al. as part of their special issue “Antimicrobial Resistance in Environmental Waters.”[138]

CHAPTER 8: PROBING THE ROBUSTNESS OF THE ENTEROTYPE HYPOTHESIS

8.1 Abstract

8.1.1 Background

The enterotype hypothesis claims that discrete patterns of microbial compositions are indicators or biomarkers of human health. Here the algorithms defining these microbial clusters are evaluated for robustness and consistency across methods.

8.1.2 Materials and Methods

The original enterotyping methods of Partitioning Around Medoids (PAM) and the Dirichlet Mixture Model (DMM) are used with several large modern datasets and their results are compared. A method similar to PAM, coseq, originally designed for clustering RNA-seq while using compositional transformations of data, is also evaluated in its enterotype prediction.

8.1.3 Results

The DMM predicts an increasing number of enterotypes as the number of samples subsampled from a cohort increases. PAM predicts various numbers of enterotypes based on the normalization scheme used. The best-suited number of clusters determined by coseq is also sensitive to the compositional transformation used.

8.1.4 Discussion

While the inconsistency of results across methods alone cannot refute the enterotype hypothesis on their own, they do suggest that the way these methods are indicated to be used in the literature are insufficient to produce results that are robust to the variety of normalizations used on microbiome data.

8.2 Background and Problem Statement

The enterotype hypothesis involves the perception of a pattern of relatively discrete clusters of different genera (enterotypes) in the human gut microbiome dominated by specific bacterial genera which give each cluster its name: *Bacteroides* (sometimes split into two groups: ET B1 and ET B2[40]), *Prevotella* (ET P) and *Ruminococcus* (ET R)[147]. These enterotypes have been suggested as a means of classification that could simplify the analysis of the complex microbial relationships in the gut environment. Enterotypes have also been proposed for use as indicators of biological trends in disease outcomes as diverse as obesity, diabetes, gut cancers and chronic gut inflammatory diseases[147–149]. However, the enterotype hypothesis remains controversial as many researchers interpret these microbial patterns as being transitions occurring on a gradient rather than as discrete units, a perspective which may be further supported by the ability of individuals to change enterotypes in some circumstances[148,150–153]. There is also contention that this enterotype description, especially the suggestive title of "driver" for the dominant taxon, is overreaching in suggesting interchangeable characterizations across datasets and that the familiar term "biomarker" would be more appropriate terminology[154].

In addition to the controversy surrounding the biological interpretations of enterotypes, increasing awareness to the sensitivity of the results of microbiome analyses to data normalization and artifacts resulting from the improper treatment of its nature of compositional data, have raised concerns that the enterotype pattern could similarly be an artifact of or dependent on the analytic pipelines employed, calling into question its robustness and ultimate utility. For example, the number of enterotypes of a more statistically rigorous method[148] like that of Dirichlet multinomial mixture (DMM) modeling are known to sometimes disagree with the partitioning around medoids (PAM) clustering algorithm used in the initial formulation of enterotypes[147,155]. Additionally, as there is no consensus agreement as to the proper normalization strategy for general microbiome analyses, the PAM method's reliance on relative abundance as its sole normalization strategy becomes suspect especially since related *k*-means clustering results are known to be susceptible to the normalization or standardization of the data[60,156]. Put another way, the enterotypes predicted by PAM may be an artifact of the sole normalization scheme used in its formulation. This concern regarding normalizations in enterotypes is especially important as researchers have used all manner of normalization in the analysis of differential abundances in, for instance, diseased versus healthy patients, and establishing a consistent strategy of normalizing for enterotype analysis and differential abundance detection may have merit. There is also some doubt as to the mathematical suitability of methods like *k*-means clustering to determine enterotypes given that the algorithm assigns samples to exclusive categories. Methods akin to generative models that permit for the discussion of hidden latent variables that explain distributional trends observed in taxonomic abundances, like the Dirichlet multinomial mixture model itself, may be more suitable to describe such data. It is also concerning that despite the controversy in the biological interpretation of enterotypes and inconclusive or limited follow-up investigations of methods characterizing enterotypes, several important proponents of the enterotype hypothesis have produced a website (http:www.enterotypes.org) for automatically converting genus-level taxonomic tables into enterotypes with limited discussion of caveats beyond a brief protocol in the form of a flowchart that essentially presupposes enterotypes. In addition to evaluating these two pioneering methods in the determination of enterotypes, we further consider a k-means based approach to clustering that works with compositional transformations appropriate to microbiome data[82].

8.3 Attributions

My role in this project was the conceptualization and experimental design, compilation of normalization methods from existing software packages, modifying and extending the DMM and PAM approaches, and evaluation of the algorithms used in the definition of enterotypes. Dr. Shan Sun helped with constructing bioinformatics pipelines for taxonomic classifications for datasets used in prototyping, but not in the final work. Dr. Fodor oversaw the work and aided in the experimental design.

8.4 Materials and Methods

8.4.1 Datasets Evaluated for Enterotypes

The human gut microbiome studies considered in this aim are detailed in [Table 8.1]. These datasets include medium and large sized Western and non-Western cohorts. The inclusion of non-Western cohorts is important because others have observed different ratios of the enterotypes in such cohorts, however, the current PAM-based enterotyping method also includes an industrialized non-Western Chinese cohort[154,157]. The restriction to the use of 16S data rather than the more detailed WGS is justified by the fact that both PAM and DMM were formulated using genus level information, which 16S sequencing can reliably provide, and the use of 16S data in the original formulation of enterotypes. The taxonomic classifications at the genus level are derived from OTU tables provided in the supplemental materials (YOC, AGP) or QIIME2 (HMP, C7K) through the Qiita database. Note that there are expected to be slight differences in the taxonomic classification between these datasets, but the enterotype hypothesis itself does not stipulate a preferred classifier. Datasets were then restricted to include only samples derived from fecal materials. The only data filtering performed in the original enterotypes paper is the removal of taxa remaining unclassified after the conversion to relative abundances, resulting in a relative abundance that no longer sums to one within each sample. DMM and coseq have no suggested or preferred filtering approaches.

Table 8.1: Datasets used in this study

Short Name	**Number of Samples**	**Reference**
Human Microbiome Project (HMP)	353 (restricted to fecal samples only)	Structure, function and diversity of the healthy human microbiome. The Human Microbiome Consortium. *Nature* **volume 486**, 207–214 (2012).
Healthy Young/Old Chinese (YOC)	1095	The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young. Bian G et al. mSphere. 2017 Sep 27;2(5).
China Single Province Cohort (C7K)	7009	Regional variation limits applications of healthy gut microbiome reference ranges and disease models. He et al. Nature Medicine volume 24, 1532–1535 (2018).
American Gut Project (AGP)	9511	American Gut: an Open Platform for Citizen Science Microbiome Research. McDonald D et al. mSystems. 2018 May 15;3(3). pii: e00031-18.

8.4.2 Distance Metrics Employed

The study that originated enterotypes used the Jensen-Shannon distance on the Partitioning Around Medoids clustering algorithm[147]. The Jensen-Shannon distance is a measure of the similarity of probability distributions, and has been used as an alternative to the Bray-Curtis dissimilarity and weighted UniFrac distance in analyzing the beta-diversity of microbiome samples. The discarded fraction of unassigned reads in the original filtering protocol is problematic because as the Jensen-Shannon distance is defined for probability distributions, but the prescribed protocol does not restore the relative abundances to a sum of one after filtering, and therefore the relative abundances being used to group similar samples is not strictly a probability. Another problem with this approach that may limit the robustness of the method is the fact that different taxonomic classifiers and databases will assign different amounts of unclassified reads, and that databases and classifiers have trended to decreased amounts of unclassified taxa as databases grow and classification methods have improved. The Jensen-Shannon distance is used with the normalized reads in keeping with the original formulation of PAM, but the normalizations are also rescaled to proper probabilities.

8.4.3 Partitioning Around Medoids (PAM)

PAM is a greedy (rather than exhaustive) algorithmic approach to k-medoids clustering, which itself is a discriminative method similar to k-means clustering, but actual data points (samples) rather than means are used to seed the clusters for which minimized distances to other points are sought. Additionally, *k*-medoids can use distances or dissimilarities apart from the Euclidean distance used in standard *k*-means clustering, and this can also act to make k-medoids more robust to outliers than k-means. Normalization strategies in advance of the PAM-based determination of enterotypes will includes a subset of normalizations detailed in [Table 2.1].

8.4.4 Clustering Evaluation Indices

The number of clusters k is selected a priori, and in the PAM approach clustering suitability is evaluated using the silhouette method or the Calinski-Harabasz index[158]. In the originating work, the silhouette method is used to validate the number of clusters predicted by the Calinski-Harabasz index. The silhouette score is the average of the scores of each sample and is bound between -1 and 1, with values near zero indicating poor clustering. The score is given by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where *a(i)* is the average intra-cluster distance and *b(i)* is the smallest average distance to any cluster in which i is not a member. The Calinski–Harabasz criterion index is also a limited heuristic, which works best under certain assumptions of cluster properties[159], and compares the ratio of between-cluster (B(k)) to within-cluster (W(k)) variation multiplied by a factor involving the dataset size (n) and number of clusters (k):

$$CH(k) = \frac{B(k)(n-k)}{W(k)(k-1)}$$

8.4.5 Dirichlet Mixture Model (DMM)

The Dirichlet mixture model uses multinomial sampling where the prior is considered to be a mixture of Dirichlet components, where these components are taken to represent enterotypes. The parameters of these components are predicted using the distribution of the data to provide evidence that such a generative model (one based on enterotypes) could generate such data. Its authors claim that this better allows for their model to fit clusters (enterotypes) that may not be even in size/frequency in the human population and that such a method appropriately penalizes complex models[155]. The DMM enterotype classification method is built on raw counts provided by the taxonomic assignment at the genus level, and so different normalization methods need not be applied to it unlike the case for methods that use data normalization. The DMM was originally formulated when microbiome datasets were small, and the algorithm scales poorly with increasing dataset size. In order to evaluate its performance on larger datasets, a subsampling strategy is employed wherein 100 permutations of larger random samples are used. These are then compared against each other and against the datasets where the whole sample size is small enough to run without subsampling. The maximum subsampled size in the larger datasets is taken to be 1200, just larger than the size of the Flemish Gut Flora Project which the DMM has been stated as being successfully employed on[16].

8.4.6 coseq

The R package coseq[82] includes functionality for investigating how the compositional data transformations influence the results of *k*-means clustering, and thus potentially may be used in the determination of enterotypes. It was not specifically implemented for use on microbiome data, but it has been evaluated for the similar problem of clustering RNA-seq data. In contrast to the DMM and PAM methods, coseq is sensitive to the selection of initial cluster seeds, and so the authors recommend estimates over 5 runs to avoid such problems. In this study, the average number and standard deviation of 5 runs are reported. Rather than using a silhouette score, coseq uses a penalization for complex models like that in the DMM methodology, but its penalty function grows as the square-root of the number of clusters and the dimensions of the data multiplied by a constant calibrated from the data. The coseq clustering strategy can choose between the standard compositional transformations (ALR, ILR and CLR), in addition to a modification of the CLR transformation called "logclr" are employed. Here the ILR transformation is based on the original Gram-Schmidt procedure on the CLR-transformed data[160]. The "logclr" transformation is similar to zero-inflated models and normalizations that provide special treatment for near-zero values, but it assigns less importance to samples with weak proportions.

$$\mathrm{logCLR}(\mathrm{x_j}) = \begin{cases} -[\ln\left(1 - \ln\left[\frac{\mathrm{x_j}}{\mathrm{g(x)}}\right]\right)]^2, & \text{if } \frac{\mathrm{x_j}}{\mathrm{g(x)}} \leq 1 \\ (\ln\left[\frac{\mathrm{x_j}}{\mathrm{g(x)}}\right])^2 & \text{otherwise} \end{cases}$$

8.5 Results

8.5.1 Dirichlet Mixture Model Results

For the HMP and YOC datasets, which are small enough to fully run through the DMM, the predicted number of enterotypes for the best model fit are 3 and 6 respectively. The selection for 6 rather than 4 in the YOC dataset is based on values the difference in score of model fit where this difference is greater than 400 compared to deciding between 3 and 4 clusters (with this difference being less than 100) in the original description of the DMM on the Twins dataset[155].

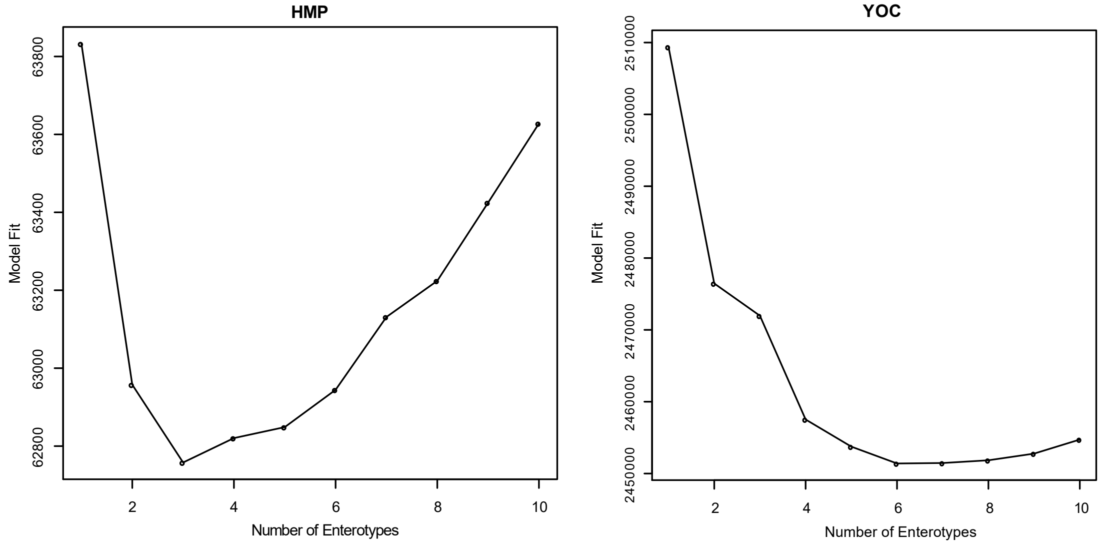

Figure 8.1: Number of enterotypes at best (minimum) model fit for HMP (left) and YOC (right) datasets

Across all datasets considered, including some of the largest currently available 16S datasets from Western and non-Western industrialized cohorts, the DMM method of determining the number of clusters or enterotypes can be sensitive to sample size using several different datasets [Figure 8.2]. In all datasets evaluated, the DMM approach predicts increasingly larger number of enterotypes with larger sample sizes. Furthermore, there is a lack of consistency between the number of enterotypes predicted between datasets, which conflicts with previous assessments of enterotypes being consistent between Western and non-Western industrialized cohorts. Here the different datasets are subsampled 100 times for each subsample size and the DMM approach is evaluated on each subsample and reports the optimal number of clusters for each subsample.

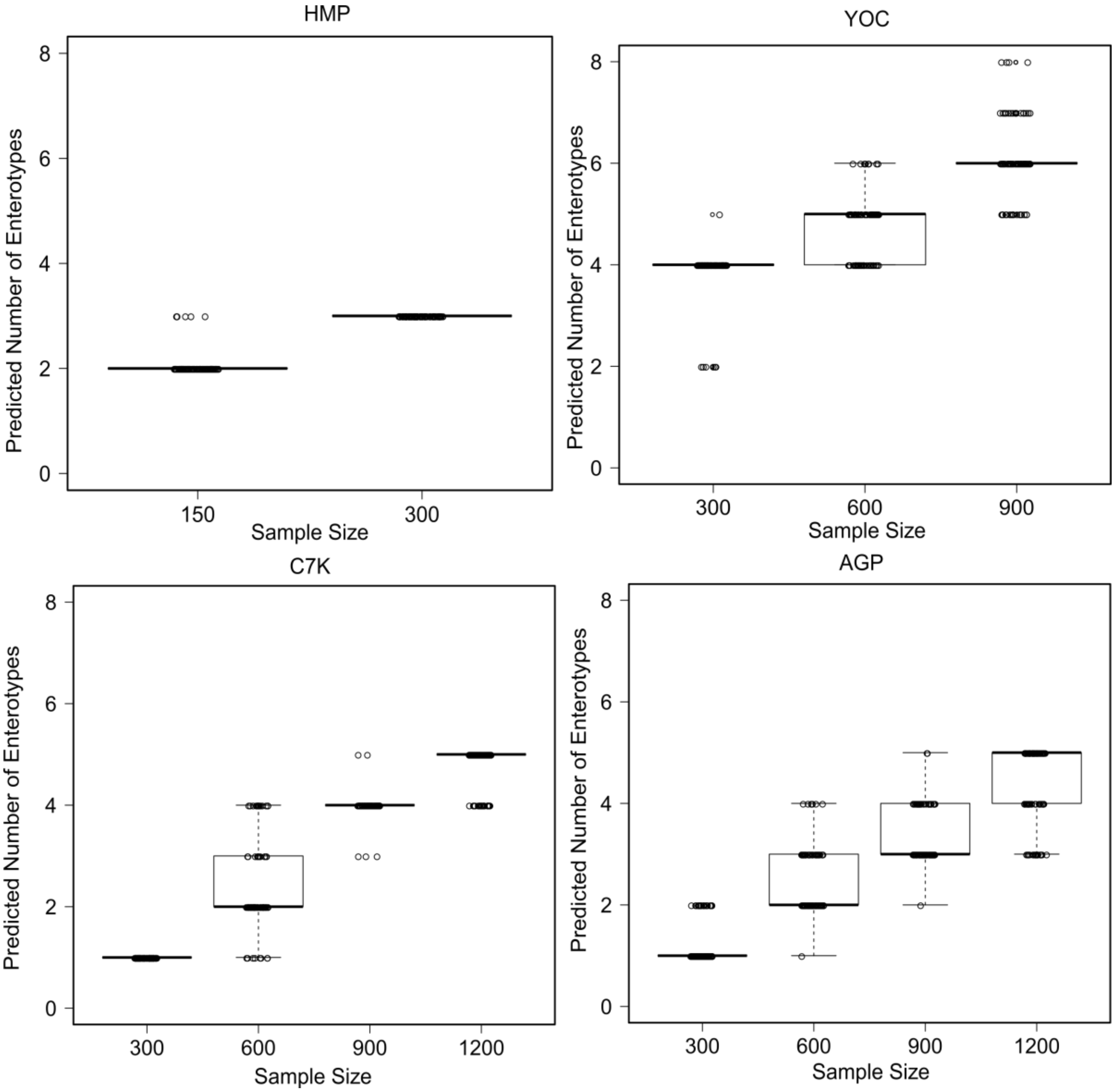

Figure 8.2: Predicted number of enterotypes grows as sample size increases in 100 permutations of random subsets of the indicated size Permutations were done without replacement for each dataset.

8.5.2 Partitioning Around Medoid Results

Similar to the inconsistencies encountered across datasets and sample sizes in the DMM, the PAM method originally used in the determination of enterotypes also predicts various optimal numbers of enterotypes subject to normalization or dataset choice. Table 8.2 indicates that the number of enterotypes depends on the normalization strategy used, but there is also some sensitivity to the dataset evaluated. Most values for HMP, YOC and AGP are near the original numbers of 2 or 3 enterotypes, but for the C7K dataset the number trends towards 4 or 5 enterotypes. Surprisingly, it is only for the relative abundance normalization used in the original formulation of enterotypes and the median ratio normalization that the number of enterotypes comes down to 2 or 3. The use of unscaled (JSD) or scaled (JSD%) data tends to not influence the number of predicted enterotypes for the HMP and YOC datasets, but there is a trend towards lower numbers of clusters as seen by the CH index in the C7K dataset for most normalizations used and higher values for the CH index within the AGP dataset. In terms of replicating the assessment validation strategy for the number of enterotypes originally employed, it is frequently the case that the two assessment methods are not consistent in predicting the same optimal number of enterotypes.

Table 8.2: The number of enterotypes predicted by PAM method depends on the normalization strategy used. SS-Silhouette Score, CH-Calinksi-Harabasz Index, JSD-Jenson-Shannon Distance, JSD%-JSD after conversion to a percentage

Normalization	HMP				YOC				C7K				AGP			
Distance	JSD		JSD%		JSD		JSD%		JSD		JSD%		JSD		JSD%	
Assessment	CH	SS	CH	SS	CH	SS	CH	SS	CH	SS	CH	SS	CH	SS	CH	SS
Raw	9	2	2	3	2	3	3	4	5	4	2	4	2	2	3	2

Relative Abundance	2	3	2	3	2	4	3	4	2	4	2	4	3	2	3	2
Log10	2	3	2	3	3	3	2	4	4	7	5	2	4	3	2	2
LogUQ	2	2	2	3	3	3	3	4	5	4	4	4	2	2	3	2
RLE	2	2	2	3	2	3	3	4	5	4	4	4	2	2	3	2
CSS	8	2	2	3	2	2	3	4	5	4	4	4	2	2	3	2
TMM	2	2	2	3	2	2	3	4	5	4	4	4	2	2	3	2
Median Ratio	2	2	2	3	2	3	3	3	2	4	4	4	2	2	3	2

Table 8.3 summarizes this data by normalization method. Unsurprisingly, the non-normalization by the use of raw data produces the highest standard deviation, but it also has the absolute highest mean value. The rescaling of normalized values to lie in the range of 0 to 1 (JSD%) leaves the mean predicted number of enterotypes for the Calinksi-Harabasz index unchanged or decreases it slightly, but the rescaling in the case of the silhouette score leads to more pronounced changes both increasing and decreasing the mean predicted number of enterotypes.

Table 8.3: Summary of number of enterotypes by normalization method

Normalization	Distance Method	CH Mean (Standard Deviation)	SS Mean (Standard Deviation)
Raw	JSD	4.50 (2.87)	2.75 (0.83)
	JSD%	2.50 (0.50)	3.25 (0.83)
Relative Abundance	JSD	2.25 (0.43)	3.25 (0.83)
	JSD%	2.50 (0.50)	3.25 (0.83)
Log10	JSD	3.25 (0.83)	4.00 (1.73)
	JSD%	2.75 (1.30)	2.75 (0.83)
LogUQ	JSD	3.00 (1.22)	2.75 (0.83)
	JSD%	3.00 (0.71)	3.25 (0.83)
RLE	JSD	2.75 (1.30)	2.75 (0.83)
	JSD%	3.00 (0.71)	3.25 (0.83)
CSS	JSD	4.25 (2.49)	2.50 (0.87)
	JSD%	3.00 (0.71)	3.25 (0.83)
TMM	JSD	2.75 (1.30)	2.75 (0.83)
	JSD%	3.00 (0.71)	3.25 (0.83)
Median Ratio	JSD	2.00 (0.00)	2.75 (0.83)

	JSD%	2.75 (0.83)	3.00 (0.71)

8.5.3 coseq Results

The coseq method, while not formally employed for enterotype determination, allows for a similar approach to that of PAM, but provides a compositional treatment of the data. The results of using different compositional transformations within the k-means clustering approach used by coseq is given in Table 8.4. Except for the ALR transformation where values run higher, the numbers of clusters/enterotypes predicted is near the 2 or 3 originally predicted and seen in Table 8.2, although the LogCLR transformation finds 3 or 4 clusters with high consistency. The ALR transformation tends to give higher variation between clustering runs, which may not be surprising given the construction of its transformation relative to a reference taxon, and this may also drive the higher number of predictions depending on the reference taxon automatically selected by the algorithm. The low variance of the clustering results of the LogCLR compared to the CLR does lend support to its authors claims of increased stability. Clustering in coseq does not yield higher enterotypes for the C7K dataset in contrast to the prediction by PAM of higher numbers of enterotypes.

Table 8.4: Mean number of clusters predicted by coseq over 5 runs with standard deviation given in parentheses

Transformation	HMP	YOC	C7K	AGP
ALR	5.8 (1.17)	5 (0.00)	5.2 (1.17)	3.4 (0.49)
ILR	2 (0.00)	4.4 (0.80)	3 (0.00)	2.6 (0.49)
CLR	2.4 (0.49)	5.2 (0.98)	3 (0.00)	2.6 (0.49)
LogCLR	3 (0.00)	6 (0.00)	4 (0.00)	4 (0.00)

8.6 Discussion

Taken together, these three methods: DMM, PAM-based clustering using different normalization strategies and coseq for compositionality-aware *k*-means clustering yield different results in terms of the number enterotypes predicted. Importantly, these differed by both normalization and dataset under consideration. The difference in performance subject to dataset choice is problematic owing to the universal claims of applicability behind the enterotype hypothesis. Data transformations that correct for compositionality do not consistently converge on the same number of enterotypes, though some methods, like the ILR and LogCLR, which have more rigorous mathematical justifications supporting their usage, show a high degree of consistency as implemented in coseq.

There are several ways to interpret the results observed for the performance of DMM. The first is that these results for the DMM method could indicate the number enterotypes could have been underestimated in the initial formulation of the enterotype hypothesis by the restrictions of their sample sizes. For both the C7K and AGP datasets there is a trend to predict a larger number of enterotypes present at the 1200 sample equivalency threshold to the Flemish Gut Flora Project data set, although this is more pronounced for the AGP data. However, this behavior is not entirely unexpected given the way that this algorithm constructs its enterotypes, but this fact has yet to be effectively communicated in the literature. Another interpretation is that the DMM is simply a flawed method of characterizing enterotypes. Yet still another interpretation is that the discrete enterotype hypothesis is false altogether and that the gradient hypothesis is correct, or discrete clusters have internal structure that these methods are not capable of deciphering. Both the gradient and sub-structure-containing enterotypes may do a better job of explaining transitions between enterotypes over time in a subject. However, it should be cautioned that this study only evaluated methods capable of producing discrete number of clusters or enterotypes. Regardless of the specifics of the interpretation of these results, it is a conservative conclusion that the existing presentation and communicated protocols for the usage of these methods are misleading in that they do not discuss these caveats mentioned herein and should be revised to reflect these observations.

While the purpose of this work was to assess these initial methods as they are currently presented in the literature and online tutorials from official sources, some comments for their improvement will be made. The DMM method is currently constrained in its detection of enterotypes by the shape of its clusters, which are essentially n-dimensional spheres of various radii. In a private communication with the corresponding author of the DMM method (Christopher Quince) has suggested that a transition to a Gaussian mixture model would allow for these n-dimensional spheres to be replaced with ellipsoids which provide more flexible cluster shapes that are more likely to capture distributional patterns of microbial compositions. This might allow multiple adjacent “enterotypes” on a curve in n-dimensional space to be condensed into a single enterotype facilitated by a bounding ellipsoid rather than a “string of pearls” of separate enterotypes. However, the reliance of the DMM on raw taxonomic counts subject to the stochastic nature of sequencing may subject the data in larger cohorts to sequencing noise related to batch effects, or other differences in experimental design if multiple experiments are being pooled. Neither of the official enterotypes methods offers suggestions to pre-filtering of data based on rarity or prevalence and such filtering could be potentially be used to adjust the sensitivity and better assess conditions that drive consistency between the two methods. It is also the case that more sophisticated scoring methods could be implemented in the existing DMM framework to allow for a hierarchical view of nested membership between enterotypes rather than one of simple same or different grouping. Another important perspective to consider is that the taxa membership that characterizes a particular enterotype may matter less than their functional competency and functional metabolism in vivo, and this true functioning of a specific microbe can be contextual to the presence/absence, perhaps at some activating threshold abundance, of other microbes. Regardless of the method selected, the biological implications for using or not using the same normalization strategy for community level assessments, such as enterotyping and ordination methods, and those for differential abundances of individual taxa remain to demonstrated in the literature.

8.7 Communication of Results

Earlier versions of this work served as the basis for a poster that was presented at The Human Microbiome Symposium at the European Molecular Biology Laboratory in Heidelberg, Germany September 16-19th, 2018. A manuscript based on these results is in preparation.

CHAPTER 9: SUMMARY

Both dysbiosis of the overarching microbial community structure and the differential abundances of specific microbes have been indicated in driving disease in humans. The importance of these microbial communities are not restricted to diseases of the gastrointestinal tract, and it has been discussed here and elsewhere that microbes play a role in the development of cancers and mental illnesses, among other forms of disease. Importantly, it is beginning to be appreciated that microbes may be a source of therapeutic molecules or be appreciated as therapeutics themselves in the form of probiotics, leading to the development of new techniques to decipher function, as well as new methods to isolate and culture useful microbes. The field itself is still growing rapidly, and now questions are changing from the conduct of microbial censuses within these various ecosystems to a functional assessment hoping to uncover mechanisms of action. It has also become a time of questioning the reproducibility and robustness of experiments conducted so far towards making reliable “gold standards” in terms of both experimental and analytical techniques. This dissertation contributes to this discussion in its exploration of various host-microbiome interactions, as well as its investigations into the analytical challenges of such explorations.

The first section of this work showed that there are only limited associations present between specific microbes and age in a non-human primate model. Diverticulosis, another age-associated condition, was investigated in a large human cohort and demonstrated limited associations between microbes and the presence or count of diverticula in the patients. The cohort size and statistical rigor of this study exceeded that of previous work on this topic which had seen significant associations. This study serves as additional evidence of issues surrounding reproducibility and the robustness of results in the microbiome literature. Moving from microbial associations with aging to the microbiota-gut-brain axis, a small patient population case study of renourishment treatment for the restoration of a healthy BMI in anorexia nervosa patients revealed that the potential microbial dysbiosis brought about by acute anorexia does not overcome the characteristic microbial signature of individual patients. However, larger future cohorts will be needed to more thoroughly explore microbial associations with renourishment treatment and to more fully understand the mechanisms behind treatment itself. Another study of the microbiota-gut-brain axis evaluated the different microbial shifts with respect to stress and demonstrated how these changes are modulated by the sex of the animal. These results provide further evidence supporting the use of mixed sex animal model cohorts, which have already been indicated by funding agencies such as the NIH, but here we highlight the case of the microbiota-gut-brain axis. The investigation of microbial communities within the scope of wastewater treatment facilities and their upstream and downstream urban waterways brings the discussion in this dissertation to a full range of *in vivo* microbial interactions. In that study, downstream communities were seen to be largely restored to that present in upstream samples, the presence of most antibiotic resistance genes were significantly decreased, but the concentrations of several antibiotics themselves remained elevated post-processing. These results contribute to the evolving discussion as to the potential role of wastewater treatment as a form of antibiotic stewardship in the face of increasing drug resistance. Finally, algorithmic methods used to define the microbial clusters supporting the controversial enterotype hypothesis were investigated as to their robustness, but numerous inconsistencies were found between different methods and normalizations.

It should be cautioned that this dissertation largely evaluated associations between microbes and their hosts and environment, but it is such studies that narrow the microbial space to be explored in follow-up experiments that work towards causal mechanisms of establishing certain microbes as beneficial or harmful. The caveats of working with compositional data have been known to the metagenomics and wider sequencing communities for some time. What is less appreciated are the variations in severity of compositionality, in part due to the distribution of relative abundances, and its potential influence on the biological interpretation of results. When these aspects are discussed, however, it is often shrouded in mathematical terminology which may fail to lead to biological insights in the ways that a comparison of actual results can. With that being said, it is also hoped that this work has been able to demonstrate the influence of compositionality on a popular hypothesis in the field of metagenomics like that of enterotypes, and can serve as a biologically-centered motivation for researchers to better address compositionality, and the robustness of results subject to different normalizations in metagenomics datasets. Even as future experimentation becomes more quantitative, much work remains to be done in assessing how those techniques extend current popular sequencing methods of exploring the microbiota that yield compositional data.

REFERENCES

1. Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164,** 337–340 (2016).

2. Turnbaugh, P. J. *et al.* The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* **449,** 804–810 (2007).

3. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

4. Hacquard, S. *et al.* Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* **17,** 603–616 (2015).

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

6. Belkaid, Y. & Hand, T. W. Role of the Microbiota in Immunity and Inflammation. *Cell* **157,** 121–141 (2014).

7. Zhang, C. *et al.* Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes. *ISME J.* **10,** 2235–2245 (2016).

8. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15,** R89 (2014).

9. Gilbert, J. A. Our unique microbial identity. *Genome Biol.* **16,** 15–17 (2015).

10. Grice, E. A. & Segre, J. A. The Human Microbiome: Our Second Genome. *Annu. Rev. Genomics Hum. Genet.* **13,** 151–170 (2012).

11. Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the archaeal and bacterial census: An update. *MBio* **7,** 1–10 (2016).

12. Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8,** 1–11 (2016).

13. Shreiner, A. B., Kao, J. Y. & Young, V. B. The Gut Microbiome in Health and in Disease. *Curr Opin Gastroenterol* **31,** 69–75 (2015).

14. Wang, B., Yao, M., Lv, L., Ling, Z. & Li, L. The Human Microbiota in Health and Disease. *Engineering* **3,** 71–82 (2017).

15. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science (80-.).* **352,** 565–569 (2016).

16. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science (80-.).* **352,** 560–564 (2016).

17. Tsilimigras, M. C. B., Fodor, A. & Jobin, C. Carcinogenesis and therapeutics: the microbiota perspective. *Nat. Microbiol.* **2,** 17008 (2017).

18. Lakritz, J. R. *et al.* Gut bacteria require neutrophils to promote mammary

tumorigenesis. *Oncotarget* **6,** 9387–96 (2015).

19. Graham, D. Y. Helicobacter pylori update: Gastric cancer, reliable therapy, and possible benefits. *Gastroenterology* **148,** 719–731.e3 (2015).

20. Xie, F.-J. Helicobacter pylori infection and esophageal cancer risk: An updated meta-analysis. *World J. Gastroenterol.* **19,** 6098 (2013).

21. Miller, A. H. & Raison, C. L. The role of inflammation in depression: from evolutionary imperative to modern treatment target. *Nat. Rev. Immunol.* **16,** 22–34 (2016).

22. Tyebji, S., Seizova, S., Hannan, A. J. & Tonkin, C. J. Toxoplasmosis: A pathway to neuropsychiatric disorders. *Neurosci. Biobehav. Rev.* **96,** 72–92 (2019).

23. Kleiman, S. C. *et al.* Daily Changes in Composition and Diversity of the Intestinal Microbiota in Patients with Anorexia Nervosa: A Series of Three Cases. *Eur. Eat. Disord. Rev.* **25,** 423–427 (2017).

24. Lyte, M. & Cryan, J. F. *Microbial Endocrinology: The Microbiota-Gut-Brain Axis in Health and Disease*. **817,** (Springer New York, 2014).

25. Aroniadis, O. C., Drossman, D. A. & Simrén, M. A Perspective on Brain-Gut Communication: The American Gastroenterology Association and American Psychosomatic Society Joint Symposium on Brain-Gut Interactions and the Intestinal Microenvironment. *Psychosom. Med.* **79,** 847–856 (2017).

26. Tsugawa, H. *et al.* Reactive oxygen species-induced autophagic degradation of helicobacter pylori CagA is specifically suppressed in cancer stem-like cells. *Cell Host Microbe* **12,** 764–777 (2012).

27. Petersen, C. & Round, J. L. Defining dysbiosis and its influence on host immunity and disease. *Cell. Microbiol.* **16,** 1024–1033 (2014).

28. Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M. & Owen, L. J. Dysbiosis of the gut microbiota in disease. *Microb. Ecol. Health Dis.* **26,** 26191 (2015).

29. Juhász, J., Kertész-Farkas, A., Szabó, D. & Pongor, S. Emergence of collective territorial defense in bacterial communities: Horizontal gene transfer can stabilize microbiomes. *PLoS One* **9,** 1–9 (2014).

30. Fan, Y., Xiao, Y., Momeni, B. & Liu, Y.-Y. Horizontal gene transfer can help maintain the equilibrium of microbial communities. *J. Theor. Biol.* **454,** 53–59 (2018).

31. Ursell, L. K., Metcalf, J. L., Parfrey, L. W. & Knight, R. Defining the human microbiome. *Nutr. Rev.* **70 Suppl 1,** S38-44 (2012).

32. Tanca, A. *et al.* Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* **5,** 79 (2017).

33. Rodgers, B., Kirley, K. & Mounsey, A. PURLs: prescribing an antibiotic? Pair it with probiotics. *J. Fam. Pract.* **62,** 148–50 (2013).

34. Hemarajata, P. & Versalovic, J. Effects of probiotics on gut microbiota: Mechanisms of intestinal immunomodulation and neuromodulation. *Therap. Adv. Gastroenterol.* **6,** 39–51 (2013).

35. McFarland, L. V. Use of probiotics to correct dysbiosis of normal microbiota following disease or disruptive events: A systematic review. *BMJ Open* **4,** (2014).

36. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8,** 1–6 (2017).

37. Tsilimigras, M. C. B. & Fodor, A. A. Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges. *Ann. Epidemiol.* **26,** 330–335 (2016).

38. Li, H. Microbiome, Metagenomics and High-Dimensional Compositional Data Analysis. *Annu. Rev. Stat. Its Appl.* 73–94 (2015).

39. Aitchison, J. *The Statistical Analysis of Compositional Data*. *Journal of the Royal Statistical Society. Series B. Methodological* **44,** (The Blackburn Press, 1982).

40. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551,** 507–511 (2017).

41. Schwager, E., Mallick, H., Ventz, S. & Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLOS Comput. Biol.* **13,** e1005852 (2017).

42. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35,** 1077–1086 (2017).

43. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5088–5090 (1977).

44. Tremblay, J. *et al.* Primer and platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* **6,** 1–15 (2015).

45. Tessler, M. *et al.* Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.* **7,** 1–14 (2017).

46. Gloor, G. B., Macklaim, J. M., Vu, M. & Fernandes, A. D. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian J. Stat.* **45,** 73 (2016).

47. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).

48. Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33,** 294–296 (2005).

49. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community

sequencing data. *Nat. Publ. Gr.* **7,** 335–336 (2010).

50. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).

51. Balvočiute, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC Genomics* **18,** 1–8 (2017).

52. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12,** 902–903 (2015).

53. 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.* **7,** 2116–25 (2013).

54. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 27 (2017).

55. Pereira, M. B., Wallroth, M., Jonsson, V. & Kristiansson, E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* **19,** 1–17 (2018).

56. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10,** 1200–2 (2013).

57. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

58. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* (2014).

59. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* **10,** (2014).

60. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 27 (2017).

61. Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* **17,** 1–14 (2016).

62. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* 1–12 (2017).

63. Zuur, A. F. *et al. Mixed Effects Models and Extensions in Ecology with R. Springer* **36,** (2009).

64. Gałecki, A. & Burzykowski, T. *Linear Mixed-Effects Models Using R*. (Springer New York, 2013).

65. Wittkowski, K. M. & Song, T. Nonparametric methods for molecular biology. *Methods Mol. Biol.* **620,** 105–153 (2010).

66. Oksanen, J. *et al.* vegan: Community Ecology Package. (2016).

67. Bray, J. & Curtis, J. T. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27,** 325–349 (1957).

68. Legendre, P. & Legendre, L. *Numerical Ecology, Volume 24*. (Elsevier, 2012).

69. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **57,** 289–300 (1995).

70. Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V. & Egozcue, J. J. It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* **26,** 322–329 (2016).

71. Leite, M. L. C. Applying compositional data methodology to nutritional epidemiology. *Stat. Methods Med. Res.* **25,** 3057–3065 (2016).

72. Pearson, K. Mathematical Contributions to the Theory of Evolution.--On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proc. R. Soc. London* **60,** 489–498 (1897).

73. Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*. (Wiley, 2015).

74. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551,** 507–511 (2017).

75. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31,** 3172–3180 (2015).

76. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Heal. Dis.* **26,** 1–7 (2015).

77. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2,** 15 (2014).

78. Fišerová, E., Donevska, S., Hron, K., Bábek, O. & Vaňkátová, K. Practical aspects of log-ratio coordinate representations in regression with compositional response. *Meas. Sci. Rev.* **16,** 235–243 (2016).

79. Andrews, S., Changizi, N. & Hamarneh, G. The isometric log-ratio transform for probabilistic multi-label anatomical shape representation. *IEEE Trans. Med. Imaging* **33,** 1890–1899 (2014).

80. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6,** 1–20 (2017).

81. Morton, J. T. *et al.* Balance Trees Reveal Microbial Niche Differentiation.

mSystems **2,** e00162-16 (2017).

82. Godichon-Baggioni, A., Maugis-Rabusseau, C. & Rau, A. Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *J. Appl. Stat.* **0,** 1–19 (2018).

83. Ahmed, T. *et al.* Calorie restriction enhances T-cell-mediated immune response in adult overweight men and women. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **64,** 1107–1113 (2009).

84. Asmuth, D. M. *et al.* Oral serum-derived bovine immunoglobulin improves duodenal immune reconstitution and absorption function in patients with HIV enteropathy. *AIDS* **27,** 2207–2217 (2013).

85. Amar, J. *et al.* Energy intake is associated with endotoxemia in apparently healthy men. *Am J Clin Nutr* **87,** 1219–1223 (2008).

86. Aguilar, M., Bhuket, T., Torres, S., Liu, B. & Wong, R. J. Prevalence of the Metabolic Syndrome in the United States, 2003-2012. *JAMA* **313,** 1973 (2015).

87. Nikolich-Žugich, J. The twilight of immunity: Emerging concepts in aging of the immune system review-article. *Nat. Immunol.* **19,** 10–19 (2018).

88. Collerton, J. *et al.* Frailty and the role of inflammation, immunosenescence and cellular ageing in the very old: Cross-sectional findings from the Newcastle 85+ Study. *Mech. Ageing Dev.* **133,** 456–466 (2012).

89. Maffei, V. J. *et al.* Biological Aging and the Human Gut Microbiota. *J. Gerontol. A. Biol. Sci. Med. Sci.* **72,** 1474–1482 (2017).

90. Mitchell, E. L. *et al.* Reduced intestinal motility, mucosal barrier function, and inflammation in aged monkeys. *J. Nutr. Health Aging* **21,** 354–361 (2016).

91. Clay, C. C. *et al.* Severe acute respiratory syndrome-coronavirus infection in aged nonhuman primates is associated with modulated pulmonary and systemic immune responses. *Immun. Ageing* **11,** 1–16 (2014).

92. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nat. Publ. Gr.* (2018). doi:10.1038/nature25979

93. Kavanagh, K. *et al.* Characterization and heritability of obesity and associated risk factors in vervet monkeys. *Obesity* **15,** 1666–1674 (2007).

94. Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci.* **110,** 9066–9071 (2013).

95. Zhang, H., Sparks, J. B., Karyala, S. V., Settlage, R. & Luo, X. M. Host adaptive immunity alters gut microbiota. *ISME J.* **9,** 770–781 (2015).

96. Schumann, R. *et al.* Structure and function of lipopolysaccharide binding protein. *Science (80-.).* **249,** 1429–1431 (1990).

97. Wilson, Q. N. *et al.* Greater Microbial Translocation and Vulnerability to Metabolic Disease in Healthy Aged Female Monkeys. *Sci. Rep.* **8,** 1–10 (2018).

98. Wei, X. *et al.* Fatty Acid Synthase Modulates Intestinal Barrier Function through Palmitoylation of Mucin 2. *Cell Host Microbe* **11,** 140–152 (2012).

99. Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci.* **110,** 9066–9071 (2013).

100. Brian West, A. The Pathology of Diverticulosis: Classical Concepts and Mucosal Changes in Diverticula. *J. Clin. Gastroenterol.* **40,** S126–S131 (2006).

101. Everhart, J. E. & Ruhl, C. E. Burden of Digestive Diseases in the United States Part I: Overall and Upper Gastrointestinal Diseases. *Gastroenterology* **136,** 376–386 (2009).

102. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12,** R60 (2011).

103. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput. Biol.* **11,** (2015).

104. Jones, R. B. *et al.* An Aberrant Microbiota is not Strongly Associated with Incidental Colonic Diverticulosis. *Sci. Rep.* **8,** 4951 (2018).

105. Barbara, G. *et al.* Gut microbiota, metabolome and immune signatures in patients with uncomplicated diverticular disease. *Gut* **66,** 1252–1261 (2017).

106. Kleiman, S. C. *et al.* The Intestinal Microbiota in Acute Anorexia Nervosa and During Renourishment. *Psychosom. Med.* 1 (2015).

107. Zipfel, S., Giel, K. E., Bulik, C. M., Hay, P. & Schmidt, U. Anorexia nervosa : aetiology , assessment , and treatment. *The Lancet Psychiatry* **0366,** (2015).

108. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci.* **102,** 11070–11075 (2005).

109. Clayton, J. A. & Collins, F. S. NIH to balance sex in cell and animal studies. *Nature* **509,** 282–283 (2014).

110. Yoon, D. Y. *et al.* Sex bias exists in basic science and translational surgical research. *Surg. (United States)* **156,** 508–516 (2014).

111. McCullough, L. D. *et al.* NIH initiative to balance sex of animals in preclinical studies: Generative questions to guide policy, implementation, and metrics. *Biol. Sex Differ.* **5,** 1–7 (2014).

112. Woitowich, N. C. & Woodruff, T. K. Implementation of the NIH Sex-Inclusion Policy: Attitudes and Opinions of Study Section Members. *J. Women's Heal.* **28,** (2018).

113. Kokras, N. & Dalla, C. Sex differences in animal models of psychiatric disorders.

Br. J. Pharmacol. **171,** 4595–4619 (2014).

114. Mclean, C. P., Asnaani, A., Litz, B. T. & G, H. S. Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. *J. Psychiatr. Res.* **45,** 1027–1035 (2011).

115. Eaton, N. R. *et al.* An Invariant Dimensional Liability Model of Gender Differences in Mental Disorder Prevalence : Evidence from a National Sample. *J. Abnorm. Psychol.* **121,** 282–288 (2012).

116. Karp, N. A. *et al.* Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat. Commun.* **8,** 15475 (2017).

117. Silva, A. F. *et al.* Sex and estrous cycle influence diazepam effects on anxiety and memory: Possible role of progesterone. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **70,** 68–76 (2016).

118. Fox, H. C., Morgan, P. T. & Sinha, R. Sex Differences in Guanfacine Effects on Drug Craving and Stress Arousal in Cocaine-Dependent Individuals. *Neuropsychopharmacology* **39,** 1527–1537 (2014).

119. Davey, K. J. *et al.* Gender-dependent consequences of chronic olanzapine in the rat: effects on body weight, inflammatory, metabolic and microbiota parameters. *Psychopharmacology (Berl).* **221,** 155–169 (2012).

120. Simpson, J., Ryan, C., Curley, A., Mulcaire, J. & Kelly, J. P. Sex differences in baseline and drug-induced behavioural responses in classical behavioural tests. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **37,** 227–236 (2012).

121. Sylvia, K. E., Jewell, C. P., Rendon, N. M., St. John, E. A. & Demas, G. E. Sex-specific modulation of the gut microbiome and behavior in Siberian hamsters. *Brain. Behav. Immun.* **60,** 51–62 (2016).

122. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555,** 623–628 (2018).

123. Marin, I. A. *et al.* Microbiota alteration is associated with the development of stress-induced despair behavior. *Sci. Rep.* **7,** 1–10 (2017).

124. Davis, D. J. *et al.* Sex-specific effects of docosahexaenoic acid (DHA) on the microbiome and behavior of socially-isolated mice. *Brain. Behav. Immun.* **59,** 38–48 (2016).

125. Coretti, L. *et al.* Sex-related alterations of gut microbiota composition in the BTBR mouse model of autism spectrum disorder. *Sci. Rep.* **7,** 1–10 (2017).

126. Sorge, R. E. *et al.* Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* **11,** 629–632 (2014).

127. Li, S. *et al.* Lachnospiraceae shift in the microbial community of mice faecal sample effects on water immersion restraint stress. *AMB Express* **7,** 1–11 (2017).

128. Asha, N. J., Tompkins, D. & Wilcox, M. H. Comparative analysis of prevalence,

risk factors, and molecular epidemiology of antibiotic-associated diarrhea due to Clostridium difficile, Clostridium perfringens, and Staphylococcus aureus. *J. Clin. Microbiol.* **44,** 2785–2791 (2006).

129. Tsilimigras, M. C. B. *et al.* Interactions between stress and sex in microbial responses within the microbiota-gut-brain axis to stress in a mouse model. *Psychosom. Med.* 1 (2018).

130. LeChevallier, M. W. & Au, K.-K. Water Treatment and Pathogen Control: Process efficiency in achieving safe drinking-water. *WHO Drink. Water Qual. Ser.* 107 (2004). doi:ISBN:1 84339 069 8

131. Kostich, M. S., Batt, A. L. & Lazorchak, J. M. Concentrations of prioritized pharmaceuticals in effluents from 50 large wastewater treatment plants in the US and implications for risk estimation. *Environ. Pollut.* **184,** 354–359 (2014).

132. Kolpin, D. W. *et al.* Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999-2000: A national reconnaissance. *Environ. Sci. Technol.* **36,** 1202–1211 (2002).

133. Manaia, C. M. *et al.* Antibiotic resistance in wastewater treatment plants: Tackling the black box. *Environ. Int.* **115,** 312–324 (2018).

134. Ventola, C. L. The antibiotic resistance crisis: part 1: causes and threats. *P T A peer-reviewed J. Formul. Manag.* **40,** 277–83 (2015).

135. Spellberg, B. The future of antibiotics. *Crit. Care* **18,** 1–7 (2014).

136. Kostyanev, T. *et al.* The Innovative Medicines Initiative's New Drugs for Bad Bugs programme: European public-private partnerships for the development of new strategies to tackle antibiotic resistance. *J. Antimicrob. Chemother.* **71,** 290–295 (2016).

137. Slater, F. R., Singer, A. C., Turner, S., Barr, J. J. & Bond, P. L. Pandemic pharmaceutical dosing effects on wastewater treatment: No adaptation of activated sludge bacteria to degrade the antiviral drug Oseltamivir (Tamiflu©) and loss of nutrient removal performance. *FEMS Microbiol. Lett.* **315,** 17–22 (2011).

138. Lambirth, K. *et al.* Microbial community composition and antibiotic resistance genes within a North Carolina Urban water system. *Water (Switzerland)* **10,** (2018).

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

140. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (2014).

141. Huttenhower, C. *et al.* High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput. Biol.* **486,** 207–214 (2015).

142. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database.

Antimicrob. Agents Chemother. **57,** 3348–3357 (2013).

143. Wang, M., Shen, W., Yan, L., Wang, X. H. & Xu, H. Stepwise impact of urban wastewater treatment on the bacterial community structure, antibiotic contents, and prevalence of antimicrobial resistance. *Environ. Pollut.* **231,** 1578–1585 (2017).

144. Yoon, Y. *et al.* Inactivation efficiency of plasmid-encoded antibiotic resistance genes during water treatment with chlorine, UV, and UV/H2O2. *Water Res.* **123,** 783–793 (2017).

145. Andersson, D. I. & Hughes, D. Evolution of antibiotic resistance at non-lethal drug concentrations. *Drug Resist. Updat.* **15,** 162–172 (2012).

146. Hughes, D. & Andersson, D. I. Selection of resistance at lethal and non-lethal antibiotic concentrations. *Curr. Opin. Microbiol.* **15,** 555–560 (2012).

147. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).

148. Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3,** (2018).

149. Knights, D. *et al.* Rethinking ‘Enterotypes’. *Cell Host Microbe* **16,** 433–437 (2014).

150. Huse, S. M., Ye, Y., Zhou, Y. & Fodor, A. A. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. *PLoS One* **7,** e34242 (2012).

151. 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.* **9,** e1002863 (2013).

152. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509,** 357–60 (2014).

153. MacDonald, N. J., Parks, D. H. & Beiko, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res.* **40,** 1–13 (2012).

154. Gorvitovskaia, A., Holmes, S. P. & Huse, S. M. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4,** 15 (2016).

155. Holmes, I., Harris, K. & Quince, C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One* **7,** e30126 (2012).

156. Mohamad, I. Bin & Usman, D. Standardization and its effects on K-means clustering algorithm. *Res. J. Appl. Sci. Eng. Technol.* **6,** 3299–3303 (2013).

157. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).

158. Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. Understanding of Internal Clustering Validation Measures. in *2010 IEEE International Conference on Data Mining* **43,**

911–916 (IEEE, 2010).

159. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. Methods* **3,** 1–27 (1974).

160. Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. Isometric Logratio Transformations for Compositional Data Analysis. *Math. Geol.* **35,** 279–300 (2003).