

STABILITY, RESISTANCE AND CHANGE IN MAMMALIAN MICROBIOTA  
AND THEIR ASSOCIATIONS WITH HOST HEALTH

by

Melanie Davis Spencer

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

2011

Approved by:

---

Dr. Anthony A. Fodor

---

Dr. Dennis R. Livesay

---

Dr. Jessica A. Schlueter

---

Dr. Todd R. Steck

©2011  
Melanie Davis Spencer  
ALL RIGHTS RESERVED

## ABSTRACT

MELANIE DAVIS SPENCER. Stability, resistance and change in mammalian microbiota and their associations with host health. (Under the direction of DR. ANTHONY A. FODOR)

What is the nature of a complex organism? Metagenomic research and its insights into biosystem function have fundamentally altered the answer to this question. High-throughput sequencing technology has revealed the multitude of microbes that live in and on human beings and other mammals. Metagenomics is beginning to uncover the relationships between microbiome and host that contribute to a complex organism's biological processes. The vast quantities of data generated by sequencing have also created analytical challenges that require new methods to identify biologically meaningful results. The research described in this dissertation applies many of these techniques to elucidate the role of microbiota in human health.

Chapter 1 presents results from our study of human choline metabolism that identified a relationship between the human gut microbiome and health. Primer design and qPCR experiments that confirm Chapter 1 results are explained in Chapter 2. Chapter 3 characterizes the microbial community from cystic fibrosis lung infection exposed to repeated courses of antibiotic therapy. An experiment designed to improve the resolution of ARISA, a metagenomic profiling technique, is described in Chapter 4. In Chapter 5, the relationship between gut microbial community composition and exercise in mice is investigated. In total, the work in this dissertation identifies several novel relationships between microbiota, host and environmental factors that may prove important in identifying underlying biological mechanisms that will improve human health.

## ACKNOWLEDGMENTS

The research described in Chapter 1 was supported by NIH grants DK055895, DK056350 and RR00046 (Dr. Steven Zeisel). Mouse models for Chapter 5, Experiments 1 & 2, were obtained as a part of NIH grant AR050085-01 (Dr. Timothy Lightfoot).

I would like to acknowledge the contributions of the laboratory of Dr. Steven Zeisel at the Nutrition Research Institute and The UNC Gillings School of Global Public Health at The University of North Carolina at Chapel Hill (choline metabolism study in Chapter 1), the laboratory of Dr. W. Joe Jones at Engencore at The University of South Carolina, Columbia, SC (454-FLX barcoded pyrosequencing in Chapters 1 and 3), the laboratory of Dr. Todd Steck, Biology Department, University of North Carolina at Charlotte (Cystic Fibrosis experimental design, sample collection and qPCR assays in Chapter 3), the laboratory of Dr. Timothy Lightfoot, previously with The University of North Carolina at Charlotte (mouse models and animal handling training in Chapter 5), the laboratory of Dr. Wei Jia, Nutrition Research Institute, University of North Carolina at Greensboro (metabolomics in Experiment 3, Chapter 5), Jon McCafferty (sampling, data collection and PCR in Experiment 1, Chapter 5, Peak Studio software in Chapter 5) and interns Larissa Kaiser (Experiment 2, Chapter 5), Maxine Mueller (Experiment 3, Chapter 5) and Nancy Deputy (Experiment 3, Chapter 5) for assistance in animal handling and data collection.

I also extend my heartfelt gratitude to Timothy J. Hamp, whose knowledge and exceptional molecular biology skills were invaluable to all experiments described in this dissertation, and to Dr. Anthony Fodor, whose computational expertise, intellectual curiosity and infectious love of learning supported, guided and inspired my research.

## INTRODUCTION

Traditional methods of researching the microbial communities that reside in or on human beings and mammals were limited to investigating those organisms that could be cultured, a small subset of those that exert a significant influence on host health.<sup>1</sup> The limited numbers of organisms that were identified using culture-based techniques focused attention on abundant or pathogenic species and, in the case of infection, how to eradicate those microbes. Additionally, the constraints of culture kept us blind to the nature of the community as a functioning ecosystem: its stability or lack thereof, the interdependence of its members and host, how it can be disrupted and what might be the outcomes of such disruptions.

Rapid advances in molecular methods, such as Automated Ribosomal Intergenic Spacer Analysis (ARISA) and high-throughput sequencing technologies, have heralded a revolution in microbial research. The study of metagenomes, the composite genomes of microbes in an environment, has been an important beneficiary of these advances and is only beginning to transform our understanding of microbial community composition and dynamics that have important ramifications for ecology and health. The outcomes from this metagenomic revolution have uncovered the depth of microbial diversity in a variety of environments<sup>2-8</sup> Research has found that human body habits harbor 100 trillion symbiotic microbes<sup>9</sup> that participate in processes as diverse as nutrient bioavailability<sup>10-11</sup> and pathogen resistance.<sup>12-13</sup> These studies and others have challenged conventional conceptions of complex organisms, placing human beings and other mammals squarely in the realm of biosystem ecology.

Next-generation sequencing technology has provided a window on the metagenomic complexity of human and mammalian biosystems.<sup>8-9</sup> With this opportunity has come an array of challenges, from the complications of experimental design to the management and analysis of the vast quantity of data produced to the mind-bending task of teasing meaningful results from data with interdependent relationships that are yet to be documented. Changing and improving sequencing techniques have necessitated the development of new data management schemes to store an overwhelming volume of sequence data, new algorithms to efficiently process data from a variety of platforms and new analytical methods to apply the expanding applications of sequencing results to biological problems. The research described in this dissertation has evolved alongside these advances in sequencing methods and makes some progress in overcoming a few of the challenges. At the same time, this work adds to our understanding of the relationship between the organism-associated microbial communities and host health.

The relevance to health outcomes of understanding humans as biosystems is most aptly demonstrated in Chapter 1, research that describes how the combination of human genetics, dietary choline deficiency and gut microbial community composition work together to define susceptibility to fatty liver development. In this collaboration with Dr. Steven Zeisel from the University of North Carolina at Chapel Hill, we characterized the gut microbiome using 454-FLX pyrosequencing of the 16S rRNA gene, a technique that describes microbial community composition by identifying inherent differences between microbes in the sequence of one gene region.<sup>14</sup> Time-series stool samples were the source of the metagenomic DNA we used to study the changing gut microbiome in 15 female subjects as they were placed on rigorously controlled, common diets where

dietary choline was manipulated.<sup>15</sup> The results identified two classes of bacteria, *Gammaproteobacteria* and *Erysipelotrichi*, whose baseline abundance combined with the haplotype of *PEMT*, a gene important for normal choline metabolism, predicted the degree to which our subjects developed fatty liver when they were on a choline-deficient diet. This study will be published in *Gastroenterology*, a journal focused on both clinical and basic research on the human digestive system.

The validation of our choline metabolism study results is described in Chapter 2. Using targeted qPCR of DNA from our subjects' baseline stool samples to quantify the abundance of Enterobacteriales, an order within the Gammaproteobacteria class, we confirmed our findings that the levels of these bacteria were critical in predicting fatty liver susceptibility in our subjects. To develop this assay, we designed novel primers that were both sensitive and specific for Enterobacteriales and used a variety of bioinformatics methods to test their efficacy before having the primer sequences synthesized and conducting the qPCR assay that confirmed our results.

The importance of the technology-driven shift from culture to sequencing is effectively illustrated by the changed view of the role of microbes in human health and disease. In Chapter 3, we examine implications of moving from an organism-focused, culture-centric understanding of cystic fibrosis infection to the recognition that the health consequences of this condition are driven by a microbial biosystem where the impact from the whole may well exceed that from any individual member.<sup>16-18</sup> In our study of one CF patient over the course of almost a year, we characterized the changing dynamics of infection, again using 16S rRNA gene pyrosequencing. Like other recent studies that have taken advantage of new sequencing technologies, this research confirmed that CF

infection is definitively polymicrobial<sup>13-14</sup> and identified many previously observed and novel CF microbial community members that are components of a rich and complex microbial mix. More importantly, our findings documented changing microbiome dynamics, as our patient experienced exacerbations in his condition and was treated with antibiotics to bring the infection under control. This study is among the first to observe the simultaneous dynamic reactions of multiple microbial community members over several cycles of infection exacerbation and antibiotic response<sup>19</sup> that are so familiar to patients living with CF.

In our CF study findings, we also encountered a specific and troublesome analytical limitation of pyrosequencing methodology: the fact that measures of microbial composition within a sample are relative. The consequence of this constraint is that comparisons across samples are inevitably subject to potential artifacts of the method. In longitudinal analysis, such as ours, a comparison between relative abundance of a microbe in one sample to that in another may not be meaningful if there are large differences in the total microbial burden between samples. In our results, we offered a potential solution to this problem by using qPCR, a method that provides quantification of microbes for each sample, to adjust the relative abundance measures provided by pyrosequencing to reflect differences between samples in overall bacterial counts.

At the beginning of our research, significant cost reductions in pyrosequencing had not yet been achieved. Consequently, we explored the alternative of using ARISA (Automated Ribosomal Intergenic Spacer Analysis), a low-cost, low-resolution DNA fingerprinting technique, to profile microbial communities from several preliminary studies we conducted. Even with decreasing sequencing costs, the expense of sequence-

based metagenomic research, where the volume of sequences required to analyze a single sample numbered in the thousands or tens of thousands, was not insignificant. This problem was particularly acute for preliminary research where testing hypotheses and gathering information were a prerequisite to funding to obtain funding for a full study. The experiment described in Chapter 4 attempted to improve the resolution of ARISA results. We reasoned that by combining the analysis from Sanger sequencing of the intergenic region between the 16S rRNA and 23S rRNA genes with microbial community profiles from ARISA on the samples, we could create a library of bacterial taxonomies that could, thereafter, be used to identify the types of bacteria associated with a specific ARISA signal. This outcome would offer a reasonable alternative to expensive sequencing. Unfortunately, the study uncovered two problems that we were unable to resolve: 1) intergenic sequence lengths are degenerate and redundant between bacteria types and 2) base pair variations in ARISA spectral signals, even between technical replicates.

Despite our inability to link ARISA signal to bacterial taxonomies, the value of low-resolution DNA profiling techniques has never been clearer. The importance of inexpensive metagenomic profiling methods that provide an assessment of similarities or differences between samples within 24-48 hours of collection is becoming critically apparent as anecdotal evidence has accumulated that sequencing conducted at different times, by different technicians or in different sequencing facilities is not always comparable.<sup>20</sup> The repercussions of this problem are monumental for metagenomic research. If samples from a study are to be compared, all samples must be prepared and sequenced together for comparisons to be valid. Furthermore, scarce resources demand

that the quality of each sample be verified before sequencing is undertaken to prevent wasting resources on inferior samples. These unresolved sequencing constraints underscore a significant role for ARISA and other fingerprinting techniques for the foreseeable future. Our early work on ARISA formed the foundation for a suite of ARISA analysis tools, *Peak Studio*, that were developed by Jon McCafferty and Robert Reid, which provide simple, flexible data management and analysis of results from DNA molecular profiling methods.

ARISA applications for metagenomic analysis of preliminary research are documented in Chapter 5. In a group of three related, preliminary studies, we extended the use of a high and low activity mouse model, developed by Dr. Timothy Lightfoot (NIH RO1-AR050085-01), to explore the effects of activity levels on the gut microbial community and vice versa. Experiment 1 investigated whether ARISA gut microbiota profiles were different in two mouse models, one with high activity levels and the other with low activity levels. In Experiment 2, a cross-over study in four C57/LJ mice, we identified significant differences in ARISA gut microbial profiles that distinguished between experimental phases of wheel running activity and constrained activity. Sampling and analysis methods were tested in two different mouse strains in Experiment 3.

This dissertation encompasses a broad spectrum of metagenomics methods and their applications to original research on mammalian microbiota. Its central theme focuses on the host-microbe relationship and how its stability, resistance and change affect mammalian biosystems. By providing a view of microbiota in different human and

model animal systems, this work advances our understanding of the influence of microbial communities on host health.

## TABLE OF CONTENTS

LIST OF TABLES	xvii
LIST OF FIGURES	xviii
LIST OF ABBREVIATIONS	xxi
CHAPTER 1: ASSOCIATION BETWEEN COMPOSITION OF THE HUMAN GASTROINTESTINAL MICROBIOME AND DEVELOPMENT OF FATTY LIVER WITH CHOLINE DEFICIENCY	1
1.1 Abstract	1
1.2 Background and significance	2
1.3 Materials and methods	4
1.3.1 Study subjects	4
1.3.2 Dietary manipulation	5
1.3.3 Clinical assessment	6
1.3.4 Fatty liver	6
1.3.5 Dietary compliance	7
1.3.6 Sample collection and preparation	7
1.3.7 ARISA sample preparation and analysis	8
1.3.8 Sequencing sample preparation and sequence data management	8
1.3.8 Statistical analysis	9
1.4 Results	10
1.4.1 Experimental design	10
1.4.2 Distinctly individual gut microbial communities	11
1.4.3 Effects of diet on the human gut microbiome	11

	xiii
1.4.4 Relationships among choline deficiency, fatty liver, gut microbes and host genotype	13
1.5 Discussion	15
CHAPTER 2: QUANTIFYING RELATIVE BACTERIAL ABUNDANCE THROUGH PRIMER DESIGN AND REAL-TIME QUANTITATIVE POLYMERASE CHAIN REACTION	26
2.1 Abstract	26
2.2 Background and significance	26
2.3 Materials and methods	27
2.3.1 Primer design	27
2.3.2 Sample preparation and DNA extraction	29
2.3.3 qPCR experiments	30
2.4 Results	30
2.5 Discussion	31
CHAPTER 3: STABILITY, RESISTANCE AND CHANGE IN THE SPUTUM MICROBIOME OF AN ADULT CYSTIC FIBROSIS PATIENT UNDERGOING REPEATED ANTIBIOTIC TREATMENT	41
3.1 Abstract	41
3.2 Background and significance	42
3.3 Materials and methods	44
3.3.1 Patient profile	44
3.3.2 Sample collection and preparation	45
3.3.3 DNA extraction	45
3.3.4 qPCR	45
3.3.5 16S PCR for 454-sequencing	46

3.3.6 Data and statistical analysis	47
3.4 Results	49
3.4.1 Experimental design	49
3.4.2 Quantitative PCR bacterial abundance	50
3.4.3 Pyrosequencing	51
3.4.4 Microbiome diversity	52
3.4.5 Microbial community composition	53
3.5 Discussion	57
CHAPTER 4: AUTOMATED RIBOSOMAL INTERGENIC SEQUENCE ANALYSIS AS AN EFFECTIVE ALTERNATIVE TO SEQUENCING FOR CHARACTERIZING COMPLEX BACTERIAL COMMUNITIES	73
4.1 Background and significance	73
4.2 Methods	75
4.2.1 Sample collection and preparation.	75
4.2.2 ARISA sample preparation and analysis	75
4.2.3 Sequencing sample preparation	76
4.2.4 Sequence assembly and taxonomy assignment	76
4.2.5 Intergenic spacer length identification	77
4.2.6 Analysis	77
4.3 Results	77
4.4 Discussion	80
4.5 Conclusion	82

CHAPTER 5: THE RELATIONSHIP BETWEEN ACTIVITY LEVELS AND THE GUT MICROBIAL COMMUNITY IN A MOUSE MODEL	88
5.1 Background and significance	88
5.2 Common methods	92
5.2.1 Animal care	92
5.2.2 Sample collection and preparation	92
5.2.3 ARISA sample preparation and analysis	93
5.3 Experiment 1: Gut metagenomics of high activity and low activity mouse strains	94
5.3.1 Experimental design	94
5.3.2 ARISA analysis	94
5.3.3 Results	94
5.4 Experiment 2: Effect of activity changes on mouse gut metagenomics	95
5.4.1 Experimental design	95
5.4.2 ARISA analysis	96
5.4.3 Statistical analysis	96
5.4.4 Results	97
5.4.5 Discussion	99
5.5 Experiment 3: Protocol testing for mouse activity studies	101
5.5.1 Experimental design	101
5.5.2 Fecal sample collection	101
5.5.3 ARISA analysis	102
5.5.4 Statistical analysis	102

	xvi
5.5.5 Blood sample collection	103
5.5.6 Metabolite extraction	103
5.5.7 Chemical derivatization	103
5.5.8 Gas chromatography time-of-flight mass spectrum (GC-TOFMS) analysis	104
5.5.9 GC-TOFMS data processing	104
5.5.10 Results	104
CHAPTER 6: CONCLUSIONS	118
REFERENCES	120
APPENDIX A: SUPPLEMENTAL MATERIALS AND METHODS FOR CHAPTER 1	134

## LIST OF TABLES

TABLE 2.1: Consensus sequences from Enterobacteriales and non-Enterobacteriales alignments.	34
TABLE 2.2: Primer pair candidates for Enterobacteriales selected for testing	35
TABLE 2.3: BLASTn performance of primer pairs against database of all unique gut microbiome sequences from our study.	36
TABLE 3.1: Clinical observation and medication stage.	60
TABLE 3.2: Total bacterial qPCR results compared to diversity.	61
TABLE 3.3: Comparison of Pseudomonas, Burkholderia and other bacteria relative abundance from pyrosequencing to abundance weighted by qPCR overall bacterial counts	62
TABLE 3.4: Comparison of average LOG10 relative and weighted abundance for genera	63
TABLE 4.1: Seqencing results from assembled contigs and quality checks	83
TABLE 4.2: Intergenic sequence lengths from contig assembly and their corresponding sequence classifications	83
TABLE 5.1: MasterMix formulation for PCR reactions of intergenic spacer region between 16S and 23S rRNA genes	107
TABLE 5.2: Differentially expressed metabolites from serum samples analyzed using mass spectrometry	107

## LIST OF FIGURES

FIGURE 1.1: Experimental design	19
FIGURE 1.2: Hierarchical clustering of gut microbiome samples	20
FIGURE 1.3: Diagrams of bacterial sequence proportions by time point at the class level	21
FIGURE 1.4: Distribution of Gammaproteobacteria abundance by time point by subject	22
FIGURE 1.5: Gammaproteobacteria and Erysipelotrichi abundance and subject genotype predict choline deficiency induced fatty liver	23
FIGURE 1.6: The <i>PEMT</i> SNP affects risk of developing choline deficiency induced fatty liver	24
FIGURE 2.1: Association of Enterobacteriales abundance at baseline (B1) with fatty liver development in post-menopausal subjects	36
FIGURE 2.2: Alignment of Enterobacteriales and non-Enterobacteriales consensus sequences	37
FIGURE 2.3: qPCR cycle graphs from qPCR experiments	38
FIGURE 2.4: Reproducibility of qPCR results	39
FIGURE 2.5: Enterobacteriales qPCR results for post-menopausal subjects compared to results from Log <sub>10</sub> relative sequence abundance	40
FIGURE 3.1: Timeline of treatment duration	64
FIGURE 3.2: qPCR bacterial counts by sample	64
FIGURE 3.3: Test of difference in means between samples in proximity to acute antibiotic treatment and samples not affected.	65
FIGURE 3.4: Log <sub>10</sub> total sequences by OTU ranked in descending order	65
FIGURE 3.5: Phylogenetic tree of OTUs	66
FIGURE 3.6: Relationships between qPCR abundance, diversity and acute antibiotic treatment	67

FIGURE 3.7: Analysis of phylogenetic distances between samples shows clustering of samples before and after acute antibiotic treatment	68
FIGURE 3.8: Proportions of Pseudomonas, Burkholderia and other genera by sample from pyrosequencing results in context of treatment regimens	68
FIGURE 3.9: Two-way clustering of samples based on numbers of sequences grouped by most detailed taxonomic classification at 80% confidence	69
FIGURE 3.10: Time series relative abundance of Pseudomonas and Burkholderia from sequencing versus qPCR bacteria counts	70
FIGURE 3.11: Associations between Pseudomonas and Burkholderia abundance	71
FIGURE 3.12: Test of mean difference in weighted abundance response of bacterial types to acute antibiotic treatment	71
FIGURE 3.13: Patterns of weighted abundance by sample for less abundant genera	72
FIGURE 4.1: Example of Automated Ribosomal Intergenic Spacer Analysis	85
FIGURE 4.2: Example of Sanger sequence contig assembly using CodonCodeAligner	86
FIGURE 4.3: ARISA spectrum from patient 30 showing two dominant peaks associated with intergenic lengths of 690 and 729	86
FIGURE 4.4: Comparison of normalized signal from ARISA and intergenic lengths from Sanger sequencing assemblies	87
FIGURE 4.5: Comparison of normalized signal from ARISA and intergenic lengths from Sanger sequencing assemblies after 10 base pair adjustment of ARISA	87
FIGURE 5.1: Weekly running distance by strain for SM/J and 129 mice	108
FIGURE 5.2: Hierarchical clustering of ARISA gut microbiota profiles from fecal samples of SM/J and 129 mice	108

FIGURE 5.3: Design for activity cross-over study (Experiment 2)	109
FIGURE 5.4: Activity levels for Phases 1 and 3 measured by daily distance run on an activity wheel	109
FIGURE 5.5: Principal Components Analysis showing differences between Phase 1 and Phase 3 activity levels	110
FIGURE 5.6: Tests for differences between mice in Phase 1 and Phase 3 activity levels as measured by distance run on an activity wheel	110
FIGURE 5.7: Tests for differences in activity levels between mice in running distance (km/day) over the course of the study	111
FIGURE 5.8: Mouse weights (gm) over the course of the experiment	111
FIGURE 5.9: Peak Studio ARISA microbiota profiles from baseline samples	112
FIGURE 5.10: Hierarchical clustering of ARISA gut microbiome profiles from mouse fecal samples	113
FIGURE 5.11: Principal Components Analysis on covariances of gut microbiome profiles from binned ARISA signals	114
FIGURE 5.12: Wilcoxon test of differences between PC1 from PCA of gut microbiome profiles for activity levels	115
FIGURE 5.13: Hierarchical clustering and Principal Components Analysis on covariances of gut microbiome profiles from ARISA binned signal of cecal, fecal and intestinal samples	116
FIGURE 5.14: Spectra from total ion chromatograms from mass spectrometry analysis of serum samples from mouse 129-5 and C57-4	117

## LIST OF ABBREVIATIONS

GI	gastrointestinal
rRNA	ribosomal ribonucleic acid
SNP	single nucleotide polymorphism
PCS	phosphatidylcholine synthase
PMT	phospholipid N-methyltransferase
NIH	National Institutes of Health
BMI	body mass index
AI	adequate intake
CPK	creatine phosphokinase
AST	aspartate aminotransferase
ALT	alanine aminotransferase
MRI	magnetic resonance imaging
LF:SF	liver fat to spleen fat ratio
DNA	deoxyribonucleic acid
ARISA	Automated Ribosomal Intergenic Spacer Analysis
PCR	Polymerase Chain Reaction
OTU	Operational Taxonomic Unit
RDP	Ribosomal Database Project
PCA	Principal Components Analysis
PC1	First Principal Component
WT	wild type
HET	heterozygous

HO	homozygous
LPS	lipopolysaccharide
qPCR	quantitative Polymerase Chain Reaction
RFU	relative fluorescence units
CT	cycle time
CF	Cystic Fibrosis
CFTR	cystic fibrosis transmembrane conductance regulator
IRB	Institutional Review Board
CV	coefficient of variation
T-RFLP	Terminal Restriction Fragment Length Polymorphism
DGGE	Denaturing Gradient Gel Electrophoresis
LH-PCR	Length Heterogeneity Polymerase Chain Reaction
CDC	Centers for Disease Control
IACUC	Institutional Animal Care and Use Committee
GC-TOFMS	Gas Chromatography Time of Flight Mass Spectrometry
TIC	total ion chromatograms

## CHAPTER 1: ASSOCIATION BETWEEN COMPOSITION OF THE HUMAN GASTROINTESTINAL MICROBIOME AND DEVELOPMENT OF FATTY LIVER WITH CHOLINE DEFICIENCY<sup>15</sup>

### 1.1 Abstract

Non-alcoholic fatty liver disease affects up to 30% of the U.S. population, but the mechanisms underlying this condition are incompletely understood. We investigated how diet standardization and choline deficiency influence the composition of the microbial community in the human gastrointestinal (GI) tract and the development of fatty liver under conditions of choline deficiency. We performed a 2-month in-patient study of 15 female subjects who were placed on well-controlled diets in which choline levels were manipulated. We used 454-FLX pyrosequencing of 16S rRNA bacterial genes to characterize microbiota in stool samples collected over the course of the study. The compositions of the GI microbial communities changed with choline levels of diets; each individual's microbiome remained distinct for the duration of the experiment, even though all subjects were fed identical diets. Variations between subjects in levels of *Gammaproteobacteria* and *Erysipelotrichi* were directly associated with changes in liver fat in each subject during choline depletion. Levels of these bacteria, change in amount of liver fat, and a single nucleotide polymorphism that affects choline were combined into a model that accurately predicted the degree to which subjects developed fatty liver on a choline-deficient diet. Host factors and GI bacteria each respond to dietary choline deficiency, although the gut microbiota remains distinct in each individual. We identified

bacterial biomarkers of fatty liver that results from choline deficiency, adding to the accumulating evidence that GI microbes have a role in metabolic disorders.

## 1.2 Background and significance

Intestinal microbes utilize nutrients and produce metabolites that influence a wide range of human phenotypes, including susceptibility to conditions such as obesity,<sup>21-22</sup> insulin resistance,<sup>11, 23</sup> metabolic syndrome,<sup>11, 23</sup> liver steatosis,<sup>11, 23</sup> Crohn's disease<sup>24</sup> and cancer.<sup>25-26</sup> Recent characterizations of the human gut microbiota have repeatedly observed that microbial communities are distinct, even among closely related individuals.<sup>27</sup> Nevertheless, the specific interactions between gut microbes, the environment and the human host that select for different microbial communities in different people are largely unknown.

Many studies have reported direct links between diet and the structure of the gut microbiome in mouse models. One recent example observed that microbiome structure rapidly shifts in response to a change from a low-fat, plant-based diet to a high-sugar, high-fat diet, modifying both the available metabolic pathways and actual gene expression.<sup>28</sup> This study and others like it have demonstrated that symbiotic microbes are key for access to and use of nutrients and energy from dietary sources<sup>21, 28-31</sup> and have identified relationships between specific gut bacteria and host metabolism.<sup>28, 31</sup> These studies have demonstrated the relationship between dietary exposure and microbial response, but extending these results from mouse models to humans is complicated by differences in gut structure and microbiome composition.<sup>28</sup>

A principal challenge in determining associations between the human gut microbiome and health is the difficulty of standardizing the diet during sampling.<sup>31</sup> Our study

overcomes this limitation by taking advantage of ongoing research that explores the effects of choline depletion in human subjects under a rigorously standardized diet in a hospital setting. Choline is an essential nutrient and a major methyl donor that supports physiological processes from normal metabolism to neurological development in the fetus.<sup>32</sup> Human beings obtain choline from two major sources: diet and endogenous production in the liver. The level of choline in the American diet varies substantially, with one study reporting a median intake of 284 mg/day compared with the recommended daily intake of 550 mg/day.<sup>33</sup> Low-choline diets have been associated with health problems in humans and in mouse models, including non-alcoholic fatty liver disease, neural tube defects, hepatic cancer and an increased risk of breast cancer.<sup>34</sup> Furthermore, common single nucleotide polymorphisms (SNPs) in several genes have been shown to affect choline production and metabolism.<sup>35-36</sup> One notable example is the gene, *PEMT*, which is important in the endogenous de novo synthesis of phosphatidylcholine. A common haplotype associated with a defective estrogen response element in *PEMT*'s promoter region disrupts this critical process.<sup>35-36</sup>

Gut bacteria can hydrolyze choline to form dimethylamine and trimethylamine.<sup>37-39</sup> Phosphatidylcholine has been identified as a component of bacterial cell walls in approximately 10% of bacteria.<sup>40</sup> To obtain phosphatidylcholine, prokaryotes utilize two pathways, the PMT pathway for endogenous biosynthesis and the PCS pathway that utilizes exogenous choline to produce phosphatidylcholine.<sup>41</sup> Different bacterial types may have none, one or both pathways, and pathway specificity is important to microbe-host interactions.<sup>42</sup>

Our recent metabolomic analysis of the effects of choline deficiency on human beings has identified metabolites, some of bacterial origin, that differentiated subjects who experienced organ dysfunction associated with choline deficiency from those who did not.<sup>43</sup> Other studies have used mouse models to determine bacterial influences on host metabolism and biochemistry,<sup>10-11, 23, 44-47</sup> and some have suggested that gut bacteria affect the bioavailability of dietary choline to the host.<sup>10-11, 39</sup> One recent study used vancomycin to chemically knock-out gut microbes in mice and observed a variety of metabolic changes, including increased choline in feces.<sup>48</sup> Another study reported on the effects from metabolic syndrome in a mouse model. These were ameliorated when antibiotics were administered to remove the gut microbiome.<sup>23</sup> Taken together, these studies suggest that the gut microbiome influences an organism's need for choline, as well as the health outcomes associated with choline deficiency. Our research seeks to further characterize this relationship.

### 1.3 Materials and methods

#### 1.3.1 Study subjects

Healthy female subjects (n=15), a subset of those enrolled in an NIH-funded study (DK055865) investigating choline metabolism, were recruited to participate in a gut metagenomic study and provided informed consent (approved by the Institutional Review Boards at the University of North Carolina at Chapel Hill [UNC] and at the University of North Carolina at Charlotte [UNCC]). Inclusion was contingent on a good state of health, a body mass index (BMI) of 18–34, and no history of hepatic, renal, or other chronic system disease determined by physical examination and standard clinical laboratory tests. Individuals eating unusual diets that would interfere with the study,

using drugs or medications known to alter liver metabolism or using choline-containing dietary supplements during the previous 3 months were excluded. Subjects were genotyped for the *PEMT* promoter SNP rs12325817 (Appendix A, Supplemental Table 1).

### 1.3.2 Dietary manipulation

After admission to the Clinical and Translational Research Center (CTRC) at UNC Hospitals, subjects were continuously supervised to assure protocol compliance. Participants were fed study diets, prepared in-house to protocol specifications.<sup>49</sup> Total food intake was adjusted to be isocaloric and to provide adequate intakes of macro- and micronutrients. (Appendix A, Supplemental Table 2)

During the baseline study phase (FIGURE 1.1), all participants were fed a conventional diet of normal foods containing 550 mg choline · 70 kg body weight · day [the current adequate intake (AI)<sup>50</sup>], 50 mg betaine · 70 kg body weight · day. After 10 days of this initial diet, subjects entered the choline-depletion phase during which they were fed a low-choline diet containing <50 mg choline · 70 kg body weight · day and 6 mg betaine · 70 kg body weight · day, as confirmed by chemical analysis of a sample of duplicate food portions.<sup>51-52</sup> Periodic measurements of urinary choline and betaine concentrations<sup>52</sup> were made to confirm dietary compliance. Subjects consumed the depletion diet until they developed organ dysfunction associated with choline deficiency or for 42 days if they did not develop dysfunction. Subjects were deemed to have organ dysfunction if they had a more than 5-fold increase in serum creatine phosphokinase (CPK) activity<sup>53</sup>; a more than 1.5-fold increase in aspartate aminotransferase (AST) or alanine aminotransferase (ALT); or an increase in liver fat content of >28% during the

choline-depletion diet and if these elevated measures resolved when dietary choline was restored.

During the depletion phase, if functional markers indicated choline deficiency associated organ dysfunction, subjects were transitioned to a high choline repletion diet, containing 850 mg choline · 70 kg body weight · day for 10 days. Subjects who did not manifest signs of organ dysfunction after 42 days of the low-choline diet were likewise advanced to the choline repletion diet for 10 days.

### 1.3.3 Clinical assessment

Blood and urine samples were taken and laboratory tests were performed on each subject at screening, on day 1, and at the end of each dietary phase, as well as every 3–4 days for the duration of the study to monitor the depletion and repletion status. These laboratory analyses (conducted at the McLendon Clinical Laboratory at UNC Hospitals; Clinical Laboratory Improvement Act and College of American Pathologists accredited) included measurements of AST, ALT, and CPK.

### 1.3.4 Fatty liver

Liver fat was measured by magnetic resonance imaging (MRI) at the beginning and end of the baseline diet, after 21 and 42 days of the low-choline diet, and at the end of the repletion period. Liver fat content was estimated by MRI with a Vision 41.5T clinicalMRsystem (Siemens, Malvern, PA) by using a modified “In and Out of Phase” procedure.<sup>54-55</sup> This approach used the differences in transverse magnetization intensity after an ultrabrief time interval (FLASH; echo time \_2.2 and 4.5 ms, flip angle\_80 °, and repetition time\_140 ms). Processing of successive FLASH MRI images with software from Siemens Medical Solutions (Malvern, PA) was used to estimate fat content.

Quantification of organ fat content was based on measurements across 5 images per subject and standardized using similarly measured images of spleen fat content to create a liver fat to spleen fat ratio (LF:SF), based on the assumption that spleen fat remained constant and could be used to normalize liver fat results (Appendix A, Supplemental Table 3). A 28% increase in liver fat from B1 to D2 time points was the predetermined threshold for indicating organ dysfunction secondary to choline deficiency. Liver fat measurements were subject to being mistimed by a couple of days because of MRI unavailability.

#### 1.3.5 Dietary compliance

Our diets were well tolerated by subjects in this study. If subjects did not tolerate the diet, their participation in the study was terminated. No side effects were observed other than those associated with the removal of choline (hepatic dysfunction and muscle damage). Symptoms typically resolved within days when dietary choline was re-introduced. All subjects who completed our study were unaware of the diet treatment they received; none felt sick or abnormal. No study-related physical illness occurred in any subject, nor were there any serious adverse events.

#### 1.3.6 Sample collection and preparation

Time-series stool samples were collected as indicated (FIGURE 1.1) based on the instructions provided to study personnel and study subjects. Samples were frozen at  $-80^{\circ}\text{C}$  and shipped on dry ice to The University of North Carolina at Charlotte where they were stored at  $-80^{\circ}\text{C}$ . Samples were thawed on ice and metagenomic DNA was extracted using QIAamp DNA Stool Mini-kit. Extracted DNA was quantified using a NanoDrop ND-1000 spectrophotometer and frozen at  $-20^{\circ}\text{C}$  to await further preparation.

### 1.3.7 ARISA sample preparation and analysis

Automated Ribosomal Intergenic Spacer Analysis<sup>56</sup> (ARISA) was used to detect microbial signatures in our samples. Using extracted DNA, the intergenic region between the 16S rRNA and the 23S rRNA genes was PCR amplified for each of two technical replicates using universal bacterial primers. (Appendix A, Supplemental Materials and Methods) Samples were run on an Applied Biosystems 3130 genetic analyzer and resulting spectra were analyzed and data vectors from all samples were compared using hierarchical clustering with custom JAVA code.

### 1.3.8 Sequencing sample preparation and sequence data management

Extracted DNA was thawed and used to PCR amplify the V1-V2 hypervariable regions of the 16S rRNA gene. The PCR products for 454 tagged sequencing were prepared with primers and thermalcycling parameters described in Fierer et al<sup>57</sup> (Appendix A, Supplemental Materials and Methods). PCR products were sent to the Environmental Genomics Core Facility at the University of South Carolina for 454-FLX pyrosequencing. The resulting ~213,000 sequences were subjected to quality control standards (Appendix A, Supplemental Materials and Methods) and sequences that did not meet quality criteria were removed from the dataset (Appendix A, Supplemental Table 4).

The remaining 194,781 sequences were assigned to Operational Taxonomic Units (OTUs) at  $\geq 97\%$  sequence similarity using the RDP-II infernal aligner and complete linkage clustering from the RDP web-based pipeline.<sup>58</sup> Sequences were assigned taxonomy using RDP classification software.<sup>59</sup> A minimum 50% threshold confidence score<sup>60</sup> was used to include sequences in each taxonomic group. Sequence counts were

analyzed using logged proportion abundance, standardized to the average sample size across all samples (Appendix A, Supplemental Materials and Methods).

### 1.3.8 Statistical analysis

General descriptive statistics used all available samples, 74 samples across 15 subjects. For statistical analysis, two subjects with an incomplete sample set (29, 04) were left out of any statistical test where a missing sample was required for the analysis.

Sequence counts were standardized to the average sample size across all samples and the resulting sequence frequencies were logged (adding 1 to each sequence to avoid logging zeros). For the calculation and an example, see Appendix A, Supplemental Materials and Methods.

Hierarchical clustering of OTUs at 97% and of ARISA binned signal intensities were conducted using Ward's method on standardized logged sequence proportions. Statistical analyses on RDP classifications were conducted at the most inclusive taxonomy for which results were identified (at the class level) to ensure that the highest possible numbers of sequences were used in comparisons (Appendix A, Supplemental Materials and Methods). All p-values were corrected for multiple comparisons using an adjusted p-value (Appendix A, Supplemental Materials and Methods).

In Figure 5C, we used Principal Components Analysis (PCA) to build a simple linear model which related gut metagenomic sequencing and LF:SF ratio changes. We conducted the PCA on the two taxa at the class level (*Gammaproteobacteria* and *Erysipelotrichi*) with the highest  $R^2$  values when regressed against LF:SF ratio change (Appendix A, Supplemental Table 5). We used the first component of the PCA as a regressor against the % change in LF:SF ratio from B1 to D2 (FIGURE 1.5C). In our

final model (FIGURE 1.5D), we included the subject genotype for the *PEMT* SNP with B1 abundance levels for *Gammaproteobacteria* and *Erysipelotrichi* to conduct the PCA. The wild-type genotype value was set to 1 and, for either heterozygous or homozygous genotype, the value was set to 2. Because such models are subject to over-fitting, we performed permutation procedures that produced a permuted p-value to assess the validity of our results (Appendix A, Supplemental Materials and Methods).

Statistical analyses were conducted using the R statistical package (R version 2.7.2 (2008-08-25)), JMP® 8.0 software for Microsoft Windows (SAS Institute) and Microsoft® Office Excel 2003.

## 1.4 Results

### 1.4.1 Experimental design

Healthy adult female subjects (n=15; Appendix A, Supplemental Table 1) were brought into the hospital and fed a proscribed experimental diet that included three dietary phases: 1) a standard research diet providing the current recommended level of choline, 2) a diet with very low choline and 3) a diet that included significant levels of choline to restore subjects' choline levels (FIGURE 1.1). Stool samples were obtained at time points reflecting dietary changes (FIGURE 1.1), and the gut microbiome was characterized from these samples. Although we recognize that the composition of mucosal-adherent microbes could differ from stool sample results<sup>61</sup>, invasive procedures required to sample that component were not advisable for this study. By sampling under controlled conditions as choline levels were manipulated, we observed the effects of dietary challenge on microbial community stability, identifying changes in bacterial taxa

that were coincident with controlled dietary manipulations and disease states in human beings.

#### 1.4.2 Distinctly individual gut microbial communities

Accumulating evidence has documented the diversity of gut microbial community composition between individuals, even those who are closely related.<sup>27</sup> If dietary differences primarily defined gut microbial community structure, placing subjects on common diets would erase some of those differences. To test the prediction that a common diet leads to a homogenous microbial community, we conducted hierarchical clustering on Operational Taxonomic Units (OTUs), groups of sequences identified as having at least 97% sequence similarity. Our samples clustered perfectly by subject ( $P = 6.78E-65$ ; Appendix A, Supplemental Materials and Methods) regardless of sampling time point (FIGURE 1.2A). A profile of each subject's microbiome obtained from ARISA, a DNA fingerprinting technology that does not involve direct sequencing (Appendix A, Supplemental Figure 1), also found nearly perfect clusters by subject (FIGURE 1.2B). The highly similar clustering patterns observed from these two distinct methods ensures that our results are not an artifact caused by sequencing errors that may be associated with pyrosequencing methodology.<sup>62</sup> These observations demonstrate that differences between individuals were maintained despite subjects being fed a common diet and that the established adult gut microbiome does not undergo wholesale, common change in response to short-term, common dietary perturbations.

#### 1.4.3 Effects of diet on the human gut microbiome

Although our subjects' microbiota remained distinct, we observed shifts in microbial community composition within subjects as their diets were altered. When bacterial

proportions were plotted for each sampling time point in each subject, changes were visually apparent (FIGURE 1.3). Bacterial classes with the highest abundance, such as *Bacteroidia* and *Clostridia*, were ubiquitous, whereas less abundant classes, such as *Fusobacteria* and *Bacilli*, were present in only a few subjects.

We measured phylogenetic differences between dietary time points using Unifrac<sup>63</sup> to assess the degree to which a phylogenetic tree for each sample differed in branch lengths from a subject's "community" tree, constructed from all sequences from that subject. In six subjects, the p-values from these comparisons revealed that the microbial community present when subjects consumed an *ad libitum* diet at baseline was more phylogenetically unique than would be expected by chance based on a threshold of  $P \leq 0.05$  (Appendix A, Supplemental Table 6), indicating that switching from an *ad libitum* diet to a balanced, standardized research diet substantially altered gut microbiome composition in a subset of our subjects.

To quantify any common impact of dietary changes on frequencies of individual taxa, we compared differences in bacteria abundance between study time points that corresponded to changes in diet (FIGURE 1.1): the *ad libitum* diet (B1) vs. the controlled research diet (B2), the *ad libitum* diet (B1) vs. the choline deficient diet (D2), the *ad libitum* diet (B1) vs. the high choline repletion diet (R1) and the choline deficient diet (D2) vs. the high choline repletion diet (R1). Using paired t-tests, we found no common change in abundance of any taxon from one time point to another after correcting for multiple comparisons (Appendix A, Supplemental Table 7). These results indicate that, while subjects experienced changes to their microbial community on exposure to a new diet (FIGURE 1.3), there is little compelling evidence for a common

pattern of change across subjects for individual types of bacteria. Nevertheless, test results before multiple comparison correction for *Gammaproteobacteria* ( $P = .006$ ) and *Betaproteobacteria* ( $P = .008$ ) from D2 to R1 indicate a potential impact either from restoring dietary choline (Appendix A, Supplemental Table 7) or possibly from removing the soy shake from the repletion diet (Appendix A, Supplemental Materials and Methods). Based on these results, we examined the differences in *Gammaproteobacteria* abundance between time points D2 and R in greater detail.

We plotted the distributions of *Gammaproteobacteria* for all time points (FIGURE 1.4). At D2, where subjects had been on an extended diet containing very low daily choline levels (50 mg), subjects exhibited a variety of *Gammaproteobacteria* abundance levels. When high dietary choline levels were restored to subjects' diets at the R1 time point, abundance levels in all but two subjects dropped to zero. This suggests that *Gammaproteobacteria* may be inhibited by very high levels of dietary choline. Although evidence for such a direct effect of choline on *Gammaproteobacteria* is not unprecedented in the literature,<sup>64</sup> further evidence will be necessary to confirm this observation at a threshold of statistical significance that survives correction for multiple tests.

#### 1.4.4 Relationships among choline deficiency, fatty liver, gut microbes and host genotype

We next examined whether each taxon's abundance levels at the B1 (baseline) time point, an "experiment-free" condition prior to dietary changes, could predict how subjects would respond to dietary choline insufficiency at D2 (end of the depletion phase).

Abundance levels of two taxa, *Gammaproteobacteria* (FIGURE 1.5A) and *Erysipelotrichi* (FIGURE 1.5B), were correlated to the percentage change in the LF:SF

ratio from B1 to D2. *Gammaproteobacteria* showed the strongest correlation, and the negative association with liver fat changes survived correction for multiple comparisons at a 5% false discovery rate ( $R^2 = .5679$ ,  $P = .00118$ , adjusted p-value .011) (Appendix A, Supplemental Table 5). These results suggest that a subject's baseline levels (*ad libitum* diet) of this taxon predict the degree of subject susceptibility to fatty liver when dietary choline is deficient.

Although both *Gammaproteobacteria* and *Erysipelotrichi* abundance levels offered some predictive power for fatty liver development in our subjects, we wished to explore whether multivariate models could better explain subjects' susceptibility. Principal Components Analysis (PCA) provides a simple method to combine *Gammaproteobacteria* and *Erysipelotrichi* B1 abundance levels as explanatory variables. A regression of the first principal component generated from this analysis against the percentage change in the LF:SF ratio from B1 to D2 improved predictive value for fatty liver susceptibility over either taxa alone (FIGURE 1.5C). To correct for the possibility of model over-fitting, we completed one million permutations of this analysis (Appendix A, Supplemental Materials and Methods), and the resulting permutation derived p-values supported the observed correlation.

To further refine our model, we borrowed insight from previous studies that demonstrated that subject genotype for the *PEMT* SNP influences susceptibility to choline deficiency induced fatty liver.<sup>35-36</sup> Results from Welch's *t* test (FIGURE 1.6A) confirmed that this SNP genotype would be an effective predictor of liver fat changes in our subjects ( $P = .0028$ ). We combined the *PEMT* SNP genotype values with B1 abundance information from *Gammaproteobacteria* and *Erysipelotrichi* (Appendix A,

Supplemental Materials and Methods) using PCA (FIGURE 1.6B). The first principal component from this analysis proved to be highly correlated with the change in LF:SF ratio (FIGURE 1.5D) for all our subjects ( $R^2 = 0.8388$ ,  $P = 2.4e-06$ ). A permutation based p-value, designed to correct for over-fitting remained highly significant ( $P = 7.0e-04$ ), demonstrating that these three factors together (FIGURE 1.5D), are a powerful predictor of the physiological effects of choline deficiency. We obtained similar results using multivariate regression (Appendix A, Supplemental Table 8).

## 1.5 Discussion

Our study demonstrates the importance of longitudinal experimental design and rigorous dietary control to identify changes in the gut microbiome that have potentially significant ramifications for human nutrition and health. We have verified previous findings<sup>9,27</sup> that gut microbial communities are distinctly individual and have demonstrated that there is little generalized convergence between subjects on a common diet over a two-month time period (FIGURE 1.2). Although gut microbiota remained characteristically individual, dietary changes did produce some effects, altering overall gut microbial composition when subjects shifted from their normal diets and changing the abundance of specific microbes when the essential nutrient, choline, was perturbed (FIGURE 1.3). The lack of statistically significant general microbial community convergence between subjects suggests that immigration of bacteria from food or the environment was not a major contributor to microbiome composition in our subjects.

Our results are presented in the context of non-alcoholic fatty liver disease estimated prevalence as high as 30% in the U.S. population<sup>65-66</sup> and observed interrelationships between choline, high fat diets, fatty liver, insulin resistance, diabetes and other

metabolic syndrome conditions. Previous literature suggests that changes associated with menopause in women, particularly when combined with relevant SNPs, may disrupt endogenous choline production, increase a subject's need for dietary choline and predispose post-menopausal women to developing fatty liver.<sup>35-36, 67-68</sup> If this model is correct, the subjects who deplete most on a choline deficient diet are those who are least efficient at endogenous choline production and are, therefore, the most dependent on obtaining needed choline from their normal diets. In Figure 1.5C, 3 subjects (33, 29 and 04) did not show the same strong correlation as the other subjects when the combination of *Gammaproteobacteria* and *Erysipelotrichi* abundance was plotted against LF:SF ratio changes. All 3 of these subjects, however, are WT for a SNP in the *PEMT* gene, which suggests that they are in the group of women better able to endogenously synthesize phosphatidylcholine.<sup>68</sup> Combining the SNP status of *PEMT* with the abundance of the two taxa produces a correlation with essentially no outliers (FIGURE 1.5D). These data support a model in which subjects with the ability to endogenously produce phosphatidylcholine are less dependent on the composition of the microbial community. While such a model will need further validation, the observation that *Gammaproteobacteria* abundance went to zero in all but two of our patients when dietary choline levels were high (FIGURE 1.4) lends further support to the assertion that members of this taxa are involved in choline-sensitive pathways that have implications for host health.

Animal models have also suggested a relationship between choline deficiency induced fatty liver and the gut microbiome. Dumas et al. described a microbiota-mediated mechanism underlying the development of fatty liver that mimicked choline

deficiency in mice fed high fat diets and was also associated with insulin resistance.<sup>10</sup>

This mechanism was explained by microbial flora that disrupt choline bioavailability to the host by converting choline to methylamines, although no specific taxa were named. Increases in *Proteobacteria*, the phylum that includes *Gammaproteobacteria*, were also observed in mice that were fed high fat diets and that exhibited increased obesity.<sup>31</sup> A study of metabolic endotoxemia, high-fat diets and obesity identified lipopolysaccharide (LPS), a phospholipid in the outer membrane of most gram-negative bacteria<sup>69</sup>, as a possible culprit in the chronic inflammation that accompanies metabolic dysfunction, insulin resistance and diabetes.<sup>47</sup> Recent work by Kudo et al. implicated gut-derived bacterial endotoxin in up-regulation of TNF-alpha, apoptosis of primary hepatocytes and development of liver injury in a murine model of non-alcoholic steatohepatitis.<sup>70</sup> Taken together, these studies provide support for the assertion that nutrient imbalance may trigger a bloom of inflammation-producing bacteria and concurrent metabolic dysfunction. The *Gammaproteobacteria* genera identified in our study, including *Klebsiella* spp., *Enterobacter* spp. and *Escherichia* spp., are known gram-negative bacteria with LPS-containing membranes.<sup>69</sup> Their combined association with fatty liver development suggests that these mechanisms may be at work in our subjects, as well.

Even though we cannot yet assign cause and effect, our results suggest that host genotype and specific members of the microbial community are important predictors of susceptibility to choline deficiency induced fatty liver disease (FIGURE 1.5D). Defining the interrelationships between these bacteria, host genotype and choline metabolism could begin to establish the biological mechanisms through which the gut microbiome influences human health. Such work could ultimately yield important insights into the

causes and risks associated with fatty liver disease and related, increasingly prevalent conditions, such as obesity, insulin resistance, diabetes and cardiovascular disease.

B1 V	B2 V	D1 V	D2 V	R1 V	R2 V
Baseline 10d	Depletion up to 42d			Repletion 10d	
550 mg/ 70 kg body weight	< 50 mg/ 70kg body weight			850 mg/ 70 kg body weight	
In-patient	In-patient			In-patient	

FIGURE 1.1: Experimental design. Participants were fed a controlled research diet that included adequate daily choline intake during the **baseline** period (red). During **depletion**, subjects were fed a diet very low in choline until they demonstrated signs of deficiency or for a maximum of 42 days (grey). The 10-day **repletion** diet was very rich in choline (green). Arrows indicate timing of stool samples.

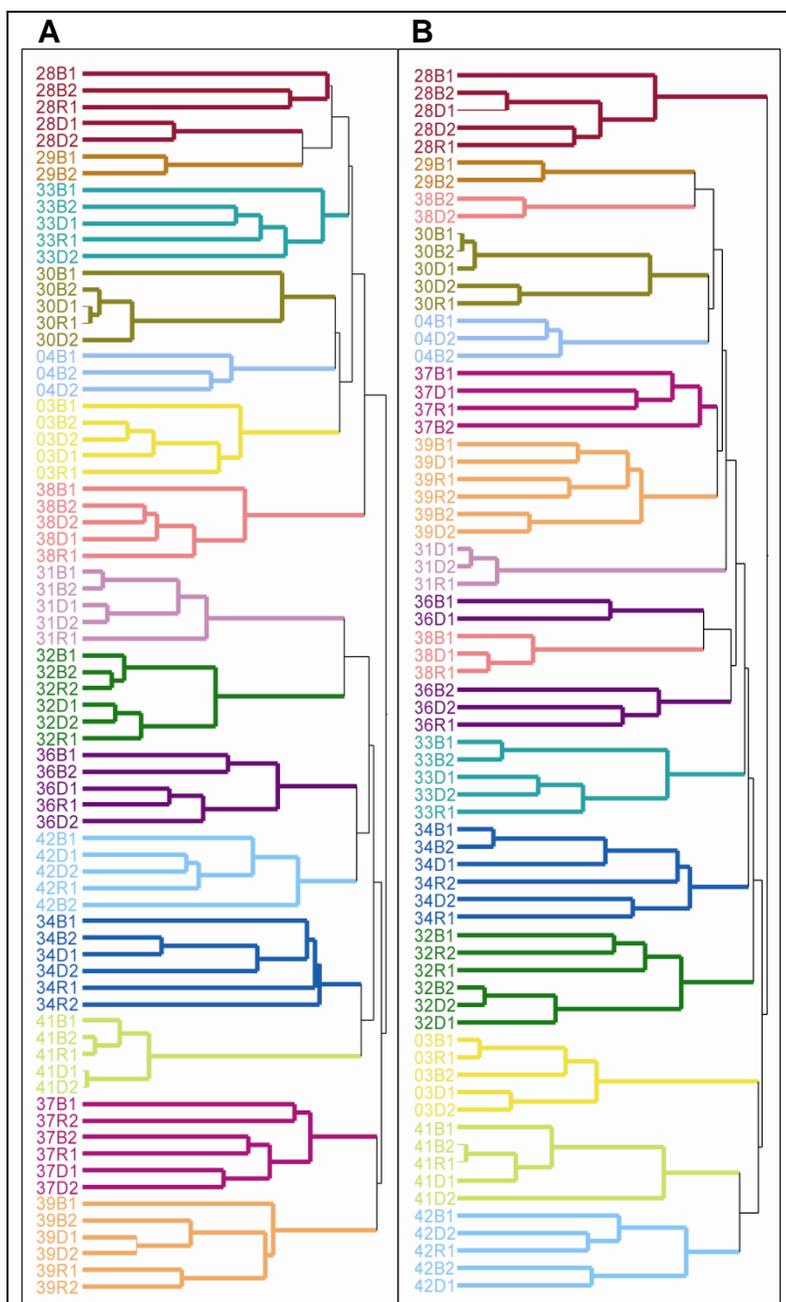


FIGURE 1.2: Hierarchical clustering of gut microbiome samples. A. Hierarchical clustering based on OTUs at 97% sequence similarity. Samples are colored by subject. B. Hierarchical clustering based on ARISA profiles, a DNA fingerprinting technique.

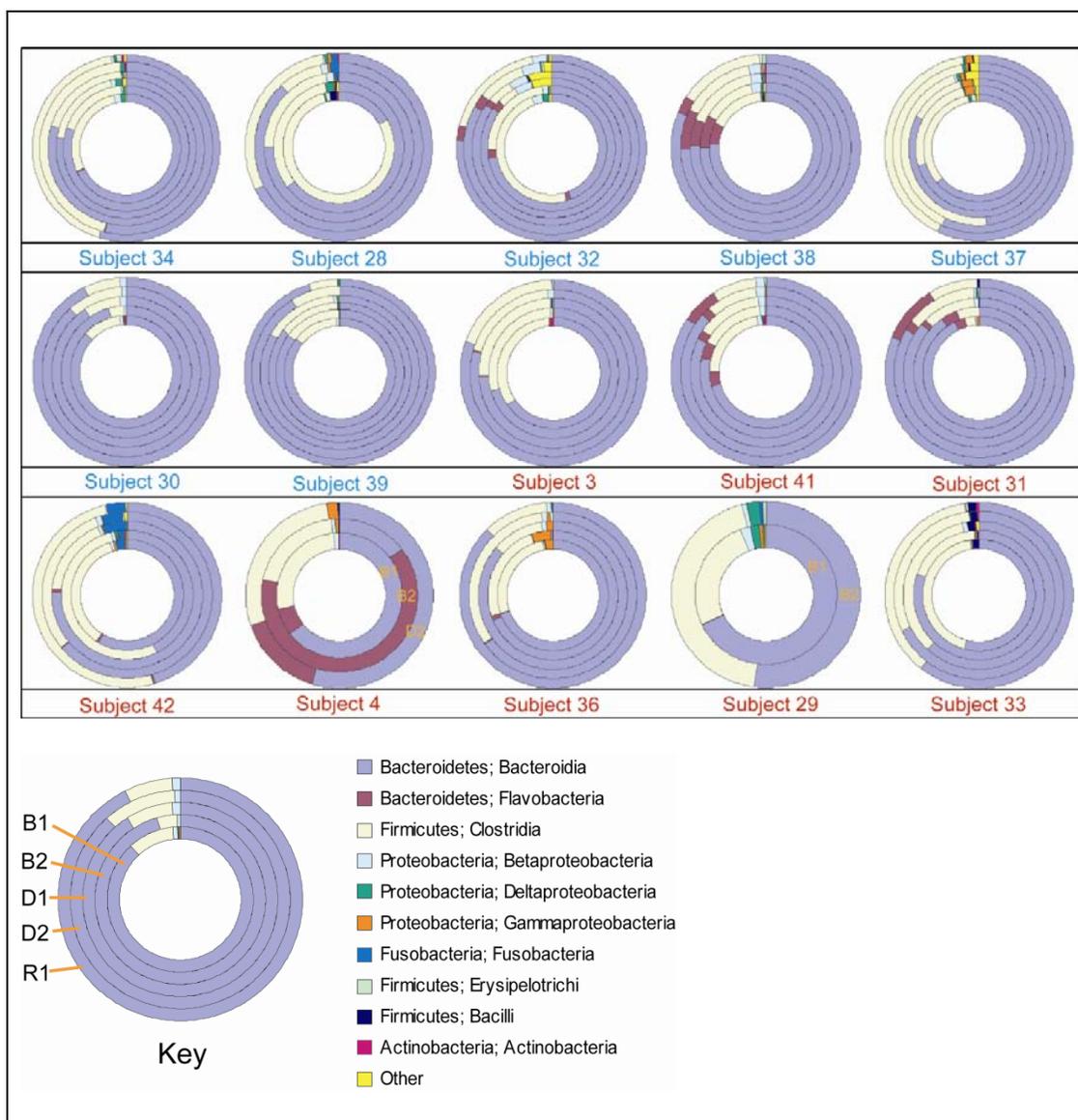


FIGURE 1.3: Diagrams of bacterial sequence proportions by time point at the class level. The key shows ordering of time points in the ring: B1-baseline, B2-standard diet, D1-choline-deficient diet, D2-end of choline-deficient diet, R1-high dietary choline and R2-second high dietary choline sample (subjects 34, 32, 37, 39). Blue subject labels designate those who developed fatty liver (liver fat change  $\geq 28\%$ ) on a choline deficient diet; those with red labels did not. Subjects 4 and 29 do not have all samples, and time points are labeled.

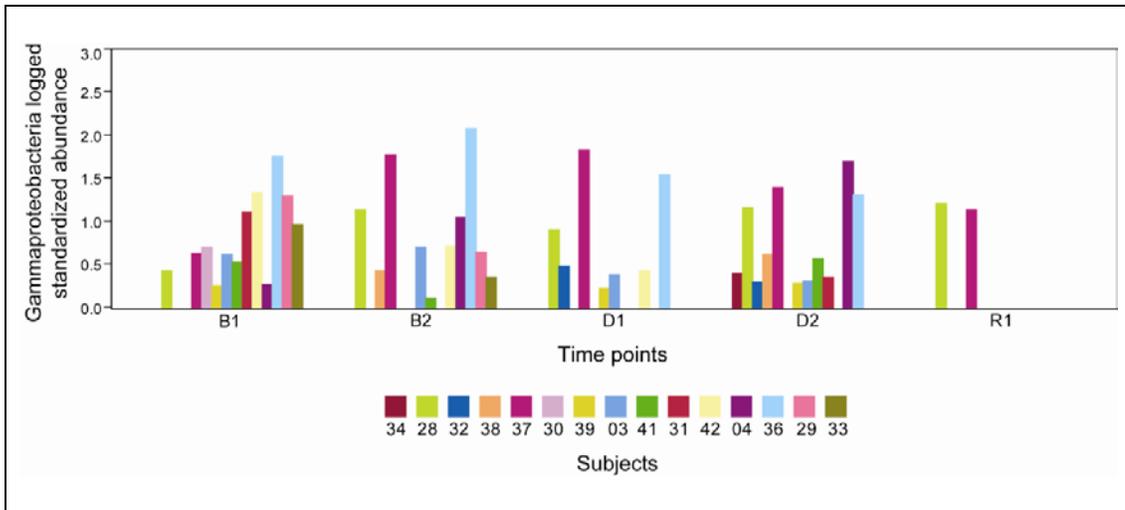


FIGURE 1.4: Distribution of Gammaproteobacteria abundance by time point by subject. Plot of logged standardized sequence frequencies ( $\log_{10}$ ) for Gammaproteobacteria colored by subject.

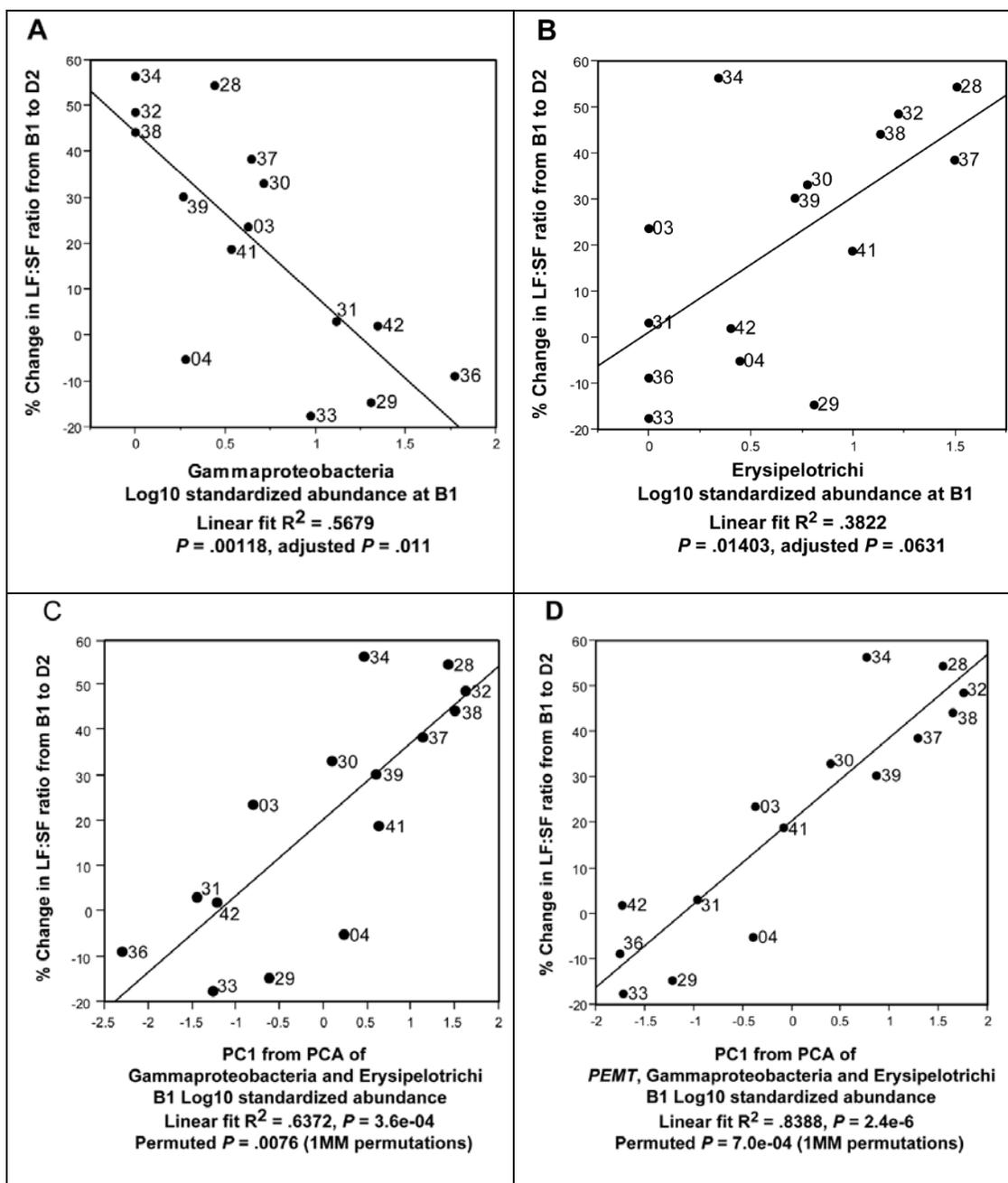


FIGURE 1.5: Gammaproteobacteria and Erysipelotrichi abundance and subject genotype predict choline deficiency induced fatty liver. Regressions of A. Gammaproteobacteria B1 abundance and B. Erysipelotrichi B1 abundance against the Liver fat/Spleen fat (LF:SF) % change from baseline (B1) to choline deficient (D2) diet. C. Regression of PCA1 from PCA of Gammaproteobacteria and Erysipelotrichi B1 abundance against the LF:SF % change from B1 to D2. D. Regression of PCA1 from PCA of *PEMT* genotype for rs12325817, Gammaproteobacteria and Erysipelotrichi B1 abundance (6B) against the LF:SF % change from B1 to D2.

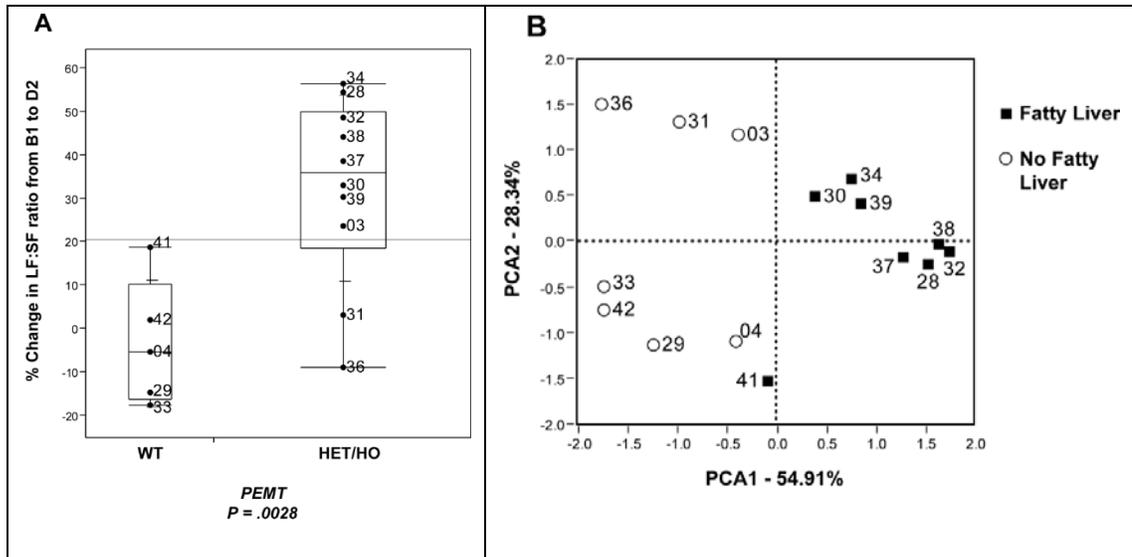


FIGURE 1.6: The *PEMT* SNP affects risk of developing choline deficiency induced fatty liver. A. Variability plot of *PEMT* genotype for SNP rs12325817 and the % change in Liver fat/Spleen fat ratio (LF:SF). WT, wildtype; HET, heterozygous; HO, homozygous. B. PCA of *PEMT* genotype for rs12325817, Gammaproteobacteria and Erysipelotrichi B1 abundance. Open circles indicate subjects who did not develop fatty liver; closed squares are subjects who did.

## CHAPTER 2: QUANTIFYING RELATIVE BACTERIAL ABUNDANCE THROUGH PRIMER DESIGN AND REAL-TIME QUANTITATIVE POLYMERASE CHAIN REACTION

### 2.1 Abstract

In our 16S rRNA gene sequencing study of the human gut microbiome (Chapter 1), we identified a linear relationship between the sequence abundance of *Gammaproteobacteria* and susceptibility to fatty liver development in subjects fed a choline deficient diet. These results required validation using another technique. We chose qPCR to test *Gammaproteobacteria* abundance and designed novel primers sensitive and specific for *Enterobacteriales*, an order within the *Gammaproteobacteria* class. Our qPCR assays confirmed the bacteria abundance we had found using 16S rRNA gene sequencing, thereby validating the relationship between health outcome and human gut microbial composition demonstrated by our study.

### 2.2 Background and significance

The strong relationship between *Gammaproteobacteria* abundance, host genotype and fatty liver development described in the previous chapter was an unexpected finding from our research. For post-menopausal subjects, the model that best explained this relationship used a more specific taxonomy, *Enterobacteriales*, an order within the *Gammaproteobacteria* class (FIGURE 2.1). We chose quantitative real time polymerase chain reaction (qPCR)<sup>71</sup> to validate the *Gammaproteobacteria* abundance in each sample and, thereby, to confirm the relationship our model had explained.

The qPCR technique can provide a quantitative assessment of DNA from a sample. The process amplifies a selected sequence, thereby allowing an otherwise undetectable amount of DNA to be detected and quantified. Primers that are both sensitive and specific to the targeted sequence are required for quantification to be accurate. If the primers have been designed effectively, they will bind to a targeted DNA sequence from the organism of interest, catalyzing the PCR chemistry and initiating amplification. For our investigation, primer sensitivity and specificity needed to span a portion of the 16S rRNA genes from the several different types of *Gammaproteobacteria* identified in our study.

## 2.3 Materials and methods

### 2.3.1 Primer design

As a first step in primer identification, we aligned the *Gammaproteobacteria* 16S rRNA sequences obtained from the samples in our study to find a common, homologous template that would recruit across the *Gammaproteobacteria* sequence dataset in a qPCR experiment. Although no effective candidate template could be identified, as an alternative, we tested *Enterobacteriales* sequences (n=608), which comprised 92% of our study's gut microbiome sequences in the *Gammaproteobacteria* class and which had exhibited the same model characteristics for our post-menopausal subjects. We used MUSCLE<sup>72</sup> and CodonCodeAligner 3.7.1 (copyright © 2009 CodonCode Corporation, Dedham, MA), to conduct multiple sequence alignment on 16S rRNA sequences from that group. From the alignment, we identified a consensus sequence template (TABLE 2.1), 203 nucleotides in length, which had high sequence similarity across all the *Enterobacteriales* sequences. To spot check for *Enterobacteriales* specificity, we

developed a database of 5000 non-*Enterobacteriales* sequences randomly selected from among 53,545 unique 16S rRNA sequences in our study. We aligned the sequences using MUSCLE and obtained a consensus sequence (TABLE 2.1) which we then compared to the *Enterobacteriales* template. The non-*Enterobacteriales* consensus sequence (gaps removed) had an overall similarity of 57.64% (Figure 2.2); however, there were regions that showed markedly higher and lower similarities, providing an indication that the *Enterobacteriales* template had areas specific enough to develop effective primer pairs.

Using the *Enterobacteriales* template, candidate primer pairs were identified using NCBI Primer-BLAST.<sup>23-24</sup> Parameters used for the primer search included:

1. PCR product size range = 75-120
2. # primers to return = 100
3. Primer melting temperatures = default
4. Specificity check organism = *Gammaproteobacteria*
5. Specificity check database = Refseq RNA
6. Primer specificity stringency = default
7. Misprimed target size deviation = default

The 20 primer pairs with the highest specificity to *Gammaproteobacteria* in the NCBI Refseq RNA database (TABLE 2.2) were selected as candidates for further testing. The primer pairs were evaluated for sensitivity and specificity for *Enterobacteriales* using BLASTn against a database of all unique sequences from our gut microbiome study (n=53,545). A strong hit in our BLAST analysis was defined as having an E-value  $\leq$  0.001. The results (TABLE 2.3) indicated that the primer pairs demonstrated a wide

range of sensitivity levels (83.53%-95.58%). However, they exhibited more consistency in specificity, with only three primer pairs recruiting sequences other than *Enterobacteriales*.

The three primer pairs with the best balance of sensitivity and specificity, PP05, PP08 and PP20 (TABLE 2.3), were chosen for further analysis. We then tested those *Enterobacteriales* sequences that were not recruited by the primer pairs to ensure that they did not all come from a single sample or taxon which, if true, could bias the qPCR assay. Taxonomic classification was conducted using the Ribosomal Database Project (RDP) sequence classification algorithm<sup>59</sup> for the *Enterobacteriales* 16S rRNA sequences that had E-values > 0.001 in each pair's BLAST results. This analysis indicated that primer pair PP05 would perform best in our qPCR experiment. A comparison of the primer to the aligned *Enterobacteriales* and non-*Enterobacteriales* consensus sequences showed that the reverse primer was critical to its performance (FIGURE 2.2). The PP05 primer pair was then sent to Integrated DNA Technologies (Coralville, IA) for synthesis.

### 2.3.2 Sample preparation and DNA extraction

Time-series stool samples from fifteen human subjects in our choline metabolism study were collected at the Clinical and Translational Research Center at the University of North Carolina at Chapel Hill (Appendix A, Supplemental Methods). Samples were frozen at -80°C and shipped on dry ice to The University of North Carolina at Charlotte where they were stored at -80°C. Samples taken at the baseline (B1) time point in the study were thawed on ice and metagenomic DNA was extracted using QIAamp DNA Stool Mini-kit (Qiagen, Valencia, CA). Extracted DNA was quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop, Wilmington, DE).

### 2.3.3 qPCR experiments

Using PicoGreen (Invitrogen, Carlsbad, CA), the relative fluorescence units (RFU) of each metagenomic DNA sample were determined and an aliquot was then normalized to a standard concentration across all samples. Quantitative PCR was performed using Quanta Biosciences PerfeCta SYBR green FastMix for iQ (Gaithersburg, MD) and run on a Bio-Rad myiQ Single Color Real-Time PCR detection system (Quanta Biosciences, Gaithersburg, MD). Each reaction was set up with 10uL of 2x FastMix, 200nM of forward PP05 and 200nM of reverse PP05 primers, 3ul standardized template and the total volume was brought to 20uL with molecular biology grade water. Samples were run at: 50C for 2 min 1x; 95C 10 min 1x; 95C 15sec; and 56C 1min 40x. Thermalcycling was immediately followed by a melting curve analysis beginning at 95C and decreasing 0.5C each cycle for 120 cycles.

## 2.4 Results

Two separate qPCR experiments (7-21-2010 and 7-30-2010) were completed on baseline (B1) samples from the fifteen choline metabolism study subjects for whom we had sequencing results and on additional subjects for whom we had samples, but not sequences. Results from qPCR are expressed in cycle time (CT), a measurement that indicates the number of cycles required for the fluorescence that is emitted in DNA binding to reach the threshold of detection. The number of cycles is logarithmically and inversely correlated to the concentration of probe-specific DNA in a sample. Therefore, higher CT values indicate lower abundance and vice versa. The immediate output from our qPCR experiments, are displayed in PCR cycle graphs (FIGURE 2.3A-B). The comparison of these sample curves with a standard curve based on serial dilutions of

known DNA quantity demonstrated that the qPCR worked properly and provided quantification of *Enterobacteriales* for each sample. To confirm the qPCR results for each sample, we checked for consistency between the two qPCR experiments by plotting the CT values from each one against the other (FIGURE 2.4).

To determine whether the qPCR confirmed our sequencing findings, we first plotted each of the B1 qPCR CT results against *Enterobacteriales* sequencing abundance levels (FIGURE 2.5A-B). The matches between qPCR and sequencing results demonstrated strong correspondence (7-21-10:  $R^2 = 0.8235$ ,  $P = 3.00E-06$ ; 7-30-10:  $R^2 = 0.8202$ ,  $P = 3.38E-06$ ) further demonstrating that qPCR validated our sequencing findings. We next checked to determine if the relationship between fatty liver development during choline deficiency and the combination of *Gammaproteobacteria* baseline abundance and PEMT host genotype and could be confirmed for *Enterobacteriales* qPCR cycle times. Because *Enterobacteriales* presence was identified as being more specific for post-menopausal subjects, we included all post-menopausal subjects for whom we had qPCR results, not just those whose samples had been sequenced, in the analysis. The combination of qPCR results and PEMT genotype from both qPCR experiments confirmed the relationship with choline deficiency induced fatty liver development we had observed previously with sequencing (FIGURE 2.5C-D).

## 2.5 Discussion

Primers of many types have been designed to target specific regions of genomes to detect presence and/or abundance of the organism or gene of interest. Commercial enterprises, such as Invitrogen and Applied Biosystems, sell primers for a variety of organisms and genes. Many published manuscripts report primer design to quantify

specific types of bacteria.<sup>23-2873</sup> Nevertheless, an extensive search of literature and internet resources indicated that the primers we created are original.

Designing primers that were both sensitive and specific for *Enterobacteriales* quantification was essential to validate our metagenomic sequencing results showing a novel and strong association between *Enterobacteriales* in our subjects and their susceptibility to fatty liver development when dietary choline was deficient. The success of our qPCR experiment solved a non-trivial technical problem and anticipated potential questions that might have arisen had our original association results been obtained using pyrosequencing as the only single method. Subsequently, we recognized the utility of these primers for additional uses.

With a non-alcoholic fatty liver disease prevalence estimated as high as 30% of adults in the United States<sup>65</sup> and an increasing incidence of associated conditions, such as obesity<sup>29-31</sup>, metabolic syndrome<sup>74-75</sup> and type 2 diabetes<sup>76</sup>, identifying factors which contribute to fatty liver development has become more urgent. In addition, non-alcoholic fatty liver disease is also associated with other serious illnesses including increased risks of liver cancer<sup>77</sup>, breast cancer<sup>78</sup>, atherosclerosis<sup>33</sup> and cardiovascular disease.<sup>79</sup> Our post-menopausal subjects who had the *PEMT* SNP, which is in the estrogen response element of the gene, were expected to be even more susceptible to choline deficiency because their lower estrogen levels would reduce gene transcription. Interestingly, higher *Enterobacteriales* abundance (both qPCR and sequence) appeared to modulate this effect, reducing the degree to which these subjects developed choline deficiency induced fatty liver. Conversely, lower *Enterobacteriales* abundance was associated with higher susceptibility. These results suggested that primers which could quantify

*Enterobacteriales* abundance, combined with measures of *PEMT* genotype, could provide ways of diagnosing individuals most likely to be affected and of tracking risk profiles over time.

TABLE 2.1: Consensus sequences from Enterobacteriales and non-Enterobacteriales alignments. Enterobacteriales (n=608) and non-Enterobacteriales (n=5000) aligned using MUSCLE. Gaps removed from Non-Enterobacteriales consensus sequence.

<b>Description</b>	<b>Sequence</b>
Enterobacteriales Consensus Sequence	CTGGTCATCCTCTCAGACCAGCTAGGGATCGTCGCC TAGGTGAGCCGTTACCCACCTACTAGCTAATCCCA TCTGGGCACATCTGATGGCATGAGGCCCGAAGGTC CCCCACTTTGGTCTTGCGACGTTATGCGGTATTAGC TACCGTTTCCAGTAGTTATCCCCCTCCATCAGGCAG TTCCACAGACATTACTCACCCGTC
Non- Enterobacteriales Consensus Sequence	TCGCGCCATCAGACGCTATCTGGACATGCTGCCTCC CGTAGGAGTTTGGACCGTGTCTCAGTTCCAATGTGG GGGACCTTCCTCTCAGAACCCCTATCCATCGTAGCC TTGGTGGGCCGTTACCCCGCCAACTAGCTAATGGAA CGCATCCCCATCGATTACCGAAATCTTTAATAATG TGACCATGCGGAATCATTATGCCATCCGGTATTAAT CTTCCTTTCGAAAGGCTATCCCCGAGTAATCGGCAG GTTGGATACGTGTTACTCACCCGTGCGCCGGTCGCC ATCAATCTATTGCAAGGRCCG

TABLE 2.2: Primer pair candidates for Enterobacteriales selected for testing.

<b>Name</b>	<b>Sequence</b>	<b>Name</b>	<b>Sequence</b>
PP01F	TCGTCGCCTAGGTGAGCCGT	PP11F	TCGTCGCCTAGGTGAGCCGT
PP01R	CGTCGCAAGACCAAAGTGGGGG	PP11R	ACCGCATAACGTCGCAAGACC
PP02F	TCGTCGCCTAGGTGAGCCGT	PP12F	AGACCAGCTAGGGATCGTCGCC
PP02R	ACGTCGCAAGACCAAAGTGGGGG	PP12R	AGTGGGGGACCTTCGGGCCT
PP03F	TCGTCGCCTAGGTGAGCCGT	PP13F	AGACCAGCTAGGGATCGTCGCC
PP03R	CGTCGCAAGACCAAAGTGGGGG	PP13R	CAAAGTGGGGGACCTTCGGGC
PP04F	TCGTCGCCTAGGTGAGCCGT	PP14F	CGTCGCCTAGGTGAGCCGTT
PP04R	ACCGCATAACGTCGCAAGACCA	PP14R	ACGTCGCAAGACCAAAGTGGGGG
PP05F	TCGTCGCCTAGGTGAGCCGTT	PP15F	AGACCAGCTAGGGATCGTCGCC
PP05R	CGTCGCAAGACCAAAGTGGGGG	PP15R	AAAGTGGGGGACCTTCGGGC
PP06F	ACCAGCTAGGGATCGTCGCC	PP16F	GACCAGCTAGGGATCGTCGCC
PP06R	CAAAGTGGGGGACCTTCGGGC	PP16R	CAAAGTGGGGGACCTTCGGGC
PP07F	ACCAGCTAGGGATCGTCGCC	PP17F	GACCAGCTAGGGATCGTCGCC
PP07R	AAAGTGGGGGACCTTCGGGC	PP17R	AAAGTGGGGGACCTTCGGGC
PP08F	TCGTCGCCTAGGTGAGCCGTT	PP18F	TCGTCGCCTAGGTGAGCCGT
PP08R	ACGTCGCAAGACCAAAGTGGGGG	PP18R	ACGTCGCAAGACCAAAGTGGG
PP09F	TCGTCGCCTAGGTGAGCCGT	PP19F	TCGTCGCCTAGGTGAGCCGTT
PP09R	AACGTCGCAAGACCAAAGTGGGGG	PP19R	CGTCGCAAGACCAAAGTGGGGG
PP10F	TCGTCGCCTAGGTGAGCCGT	PP20F	TCGTCGCCTAGGTGAGCCGTT
PP10R	CCGCATAACGTCGCAAGACCA	PP20R	ACCGCATAACGTCGCAAGACCA

TABLE 2.3: BLASTn performance of primer pairs against database of all unique gut microbiome sequences from our study. Strong hits had an E-value  $\leq 0.001$ . Highlighted primer pairs demonstrated high sensitivity and specificity and were selected for further testing.

	PP01	PP02	PP03	PP04	PP05	PP06	PP07	PP08	PP09	PP10
<b>Strong hits</b>	232	231	228	231	238	209	208	237	231	231
<b>Enterobacteriales hits</b>	232	231	228	231	238	209	208	237	231	231
<b>Sensitivity</b>	93%	93%	92%	93%	96%	84%	84%	95%	93%	93%
<b>Other bacteria hits</b>	0	0	0	0	0	0	0	0	0	0
<b>Missing Enterobacteriales</b>	17	18	21	18	11	30	31	12	18	18
<b>Specificity</b>	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	PP11	PP12	PP13	PP14	PP15	PP16	PP17	PP18	PP19	PP20
<b>Strong hits</b>	231	240	242	235	241	211	210	229	234	237
<b>Enterobacteriales hits</b>	231	210	212	235	211	211	210	229	234	237
<b>Sensitivity</b>	93%	84%	85%	94%	85%	85%	84%	92%	94%	95%
<b>Other bacteria hits</b>	0	30	30	0	30	0	0	0	0	0
<b>Missing Enterobacteriales</b>	18	9	7	14	8	38	39	20	15	12
<b>Specificity</b>	100%	88%	88%	100%	88%	100%	100%	100%	100%	100%
<b>Total Enterobacteriales sequences</b>										608
<b>Total Enterobacteriales unique sequences</b>										249

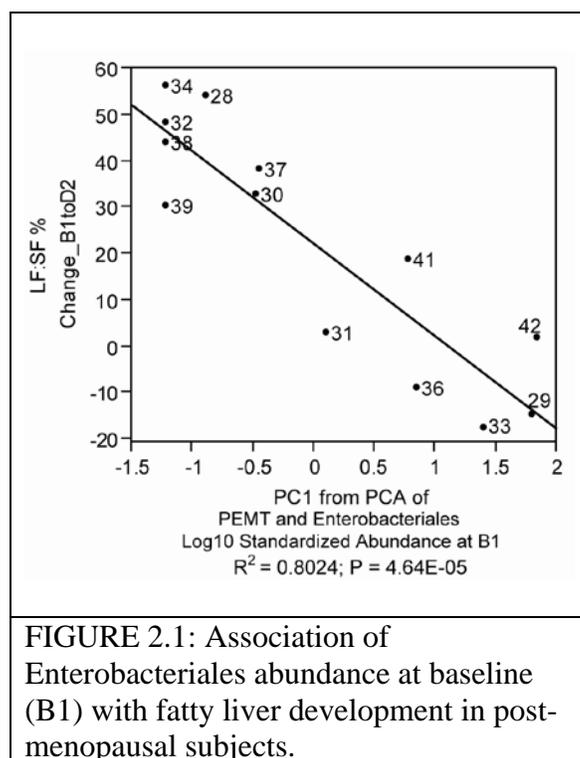


FIGURE 2.1: Association of Enterobacteriales abundance at baseline (B1) with fatty liver development in post-menopausal subjects.

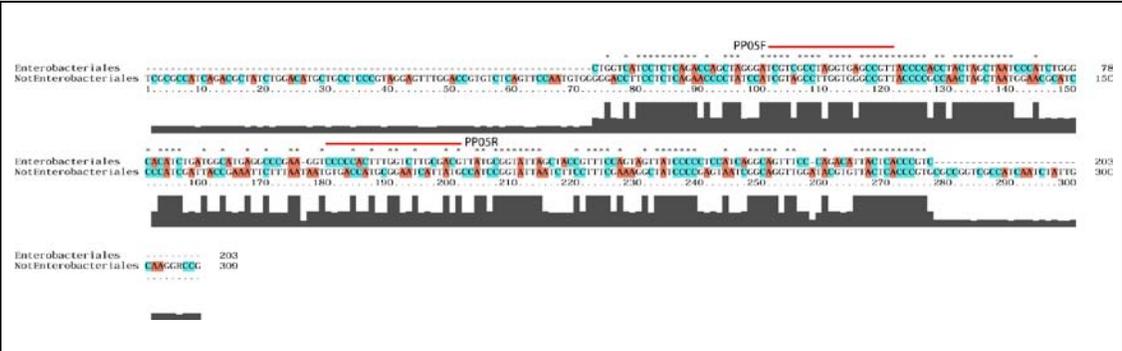
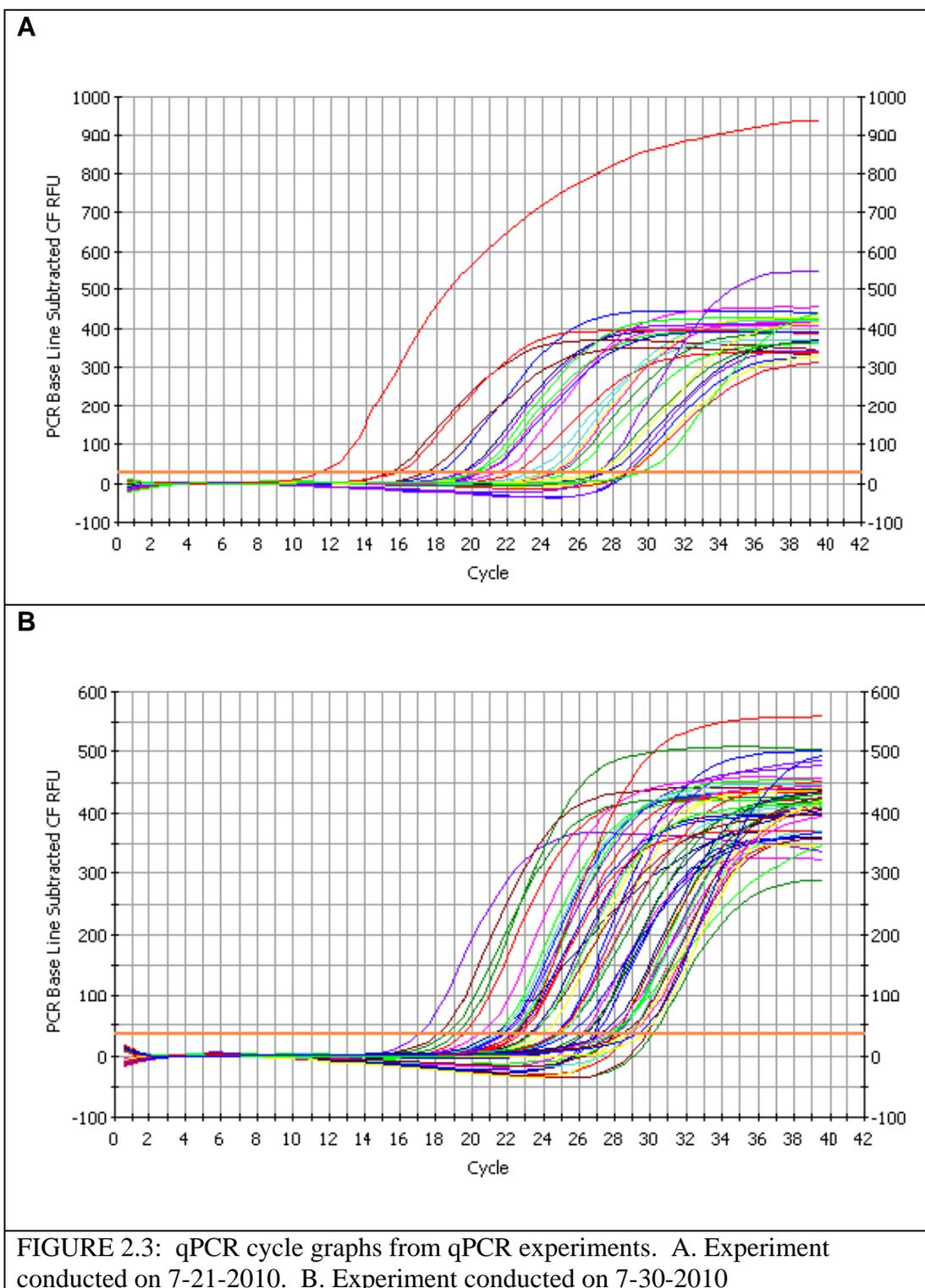


FIGURE 2.2: Alignment of Enterobacteriales and non-Enterobacteriales consensus sequences. Red lines above the alignment indicate the binding locations of the PP05 forward and reverse primers, highlighting the specificity of the reverse primer for only Enterobacteriales sequence.



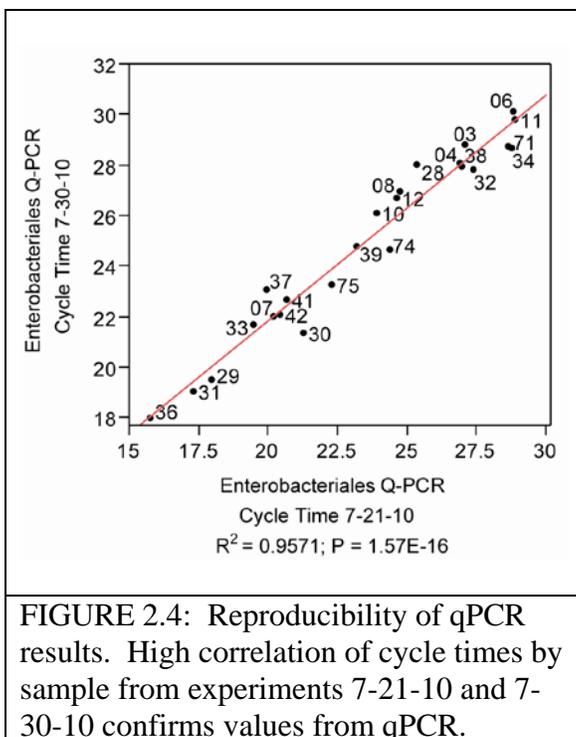


FIGURE 2.4: Reproducibility of qPCR results. High correlation of cycle times by sample from experiments 7-21-10 and 7-30-10 confirms values from qPCR.

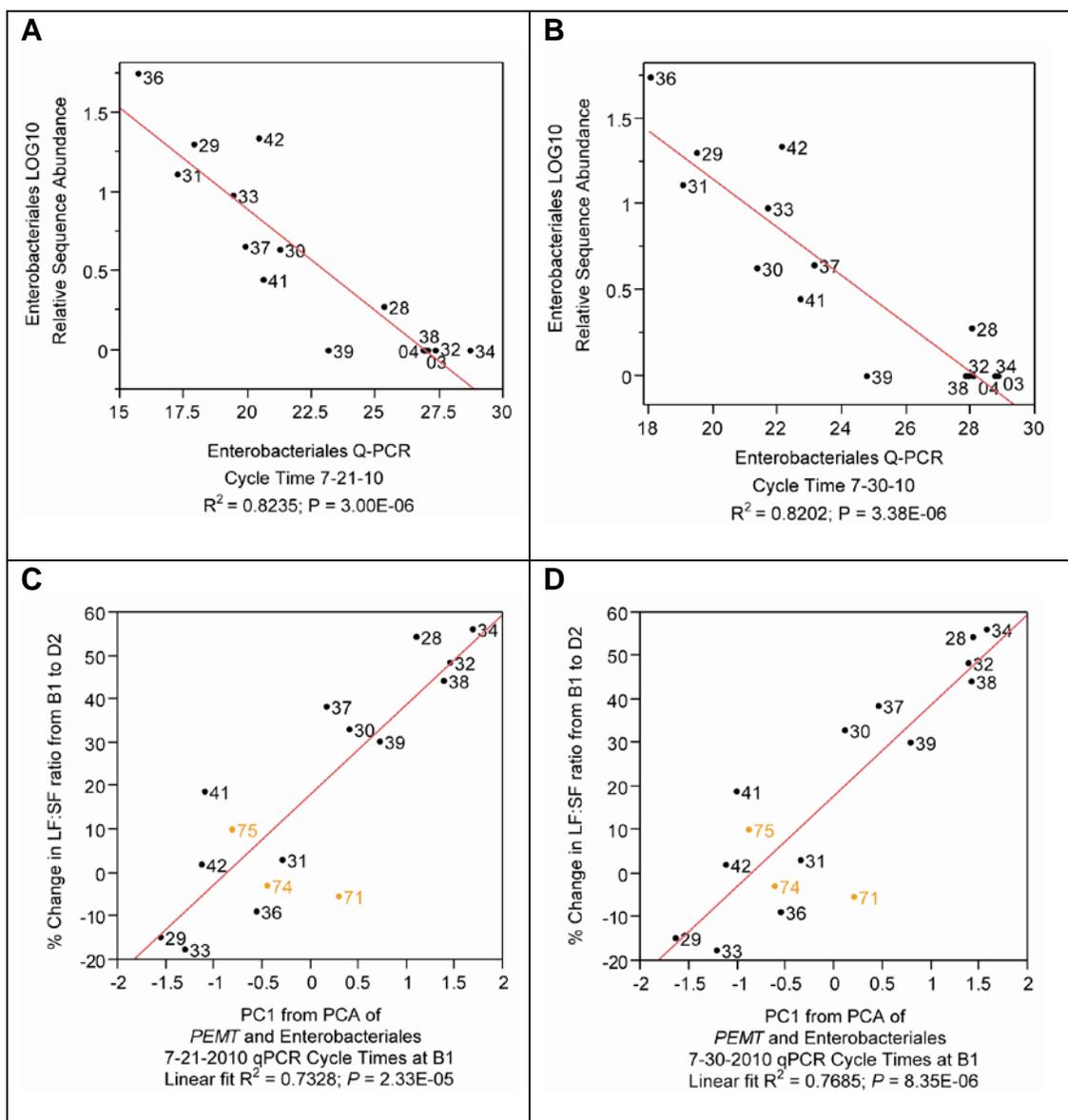


FIGURE 2.5: Enterobacteriales qPCR results for post-menopausal subjects compared to results from Log10 relative sequence abundance. Enterobacteriales qPCR plotted against Enterobacteriales sequence abundance for each subject for A. 7-21-2010 and B. 7-30-2010. Choline-deficiency induced fatty liver development (change from B1 to D2) regression against PC1 of PCA model of subject genotype for *PEMT* and Enterobacteriales qPCR from C. 7-21-2010 qPCR and D. 7-30-2010 qPCR. Subjects in orange have only qPCR and no sequencing results.

## CHAPTER 3: RESPONSE AND RESILIENCE IN THE SPUTUM MICROBIOME OF AN ADULT CYSTIC FIBROSIS PATIENT UNDERGOING REPEATED ANTIBIOTIC TREATMENT

### 3.1 Abstract

Although cystic fibrosis is caused by a genetic mutation, the manifestation of disease results from microbial colonization of the lungs that requires regular treatment, both with prophylactic therapies and with systemic antibiotics to relieve periodic acute infection. Diagnosis of infection and selection of treatments has traditionally been determined using bacterial culture and, consequently, has focused on the most abundant or recognizable pathogens as the infection source. The application of metagenomic techniques, such as pyrosequencing, to characterize the cystic fibrosis microbial community is relatively new and has revealed that lung colonization is decidedly polymicrobial.

Our study profiles cystic fibrosis infection in one adult patient over a 276-day period using 21 sputum samples collected before, during and after antibiotic treatment. It is one of the few longitudinal studies that use 16S rRNA gene pyrosequencing to characterize the cystic fibrosis lung microbiome.<sup>80-82</sup> Although the dominant pathogens in all our samples were *Pseudomonas* and *Burkholderia*, which is common in adult patients, sequencing also revealed a rich community with members from at least six different phyla and from several genera rarely identified in such infections. Our sampling scheme also allowed us to compare microbial community composition between conditions and over time, identifying a strong microbiome response to repeated antibiotic treatment that

was inevitably followed by quick community recovery. While the findings from one patient cannot be generalized broadly, our results illustrate that new techniques can yield insights into well-studied and presumably well-understood problems and suggest that metagenomic techniques and longitudinal sampling could provide insight into the clinical management of this chronic disease.

### 3.2 Background and significance

Cystic fibrosis (CF) affects approximately 30,000 people in the United States and 70,000 worldwide<sup>83</sup> and is the most common inherited chronic disease among Caucasians.<sup>84</sup> Several different mutations in the *CFTR* gene affect epithelial chloride channels and result in airway dehydration<sup>85</sup> that prevents normal airway clearance.<sup>86-87</sup> CF patients produce abnormally thick and sticky mucus in their airways that clogs the lungs and encourages chronic bacterial infections that incite a strong inflammatory response<sup>88</sup>, scarring and destroying lung tissue.<sup>89</sup> The disease can also disrupt normal pancreatic function<sup>88</sup>, preventing proper digestion and absorption of foods and resulting in bowel problems, weight loss and poor growth.<sup>83</sup> Recent advances using systemic and inhaled antibiotic treatments have extended US patients' life spans to a median survival age of 37.<sup>83</sup> Nevertheless, typical disease progression exhibits increasingly frequent and acute infection episodes, that become less responsive to treatment and substantially deteriorate quality of life as patients age.

Pulmonary infections in CF patients have been extensively researched using bacterial cultures and were traditionally characterized as having limited diversity<sup>16</sup>; however, the culture-independent studies have revealed a complex microbiota containing both aerobes and anaerobes.<sup>81-82, 90-92</sup> Clinical and study data have suggested a common succession

pattern of pathogens over the course of the disease. In early childhood, *Haemophilus influenzae* and *Staphylococcus aureus*, typically found in healthy airways, dominate the microbiome.<sup>93</sup> These are replaced by the microbes uncommonly associated with human airways, such as *Pseudomonas aeruginosa*, *Stenotrophomonas maltophilia* and/or *Burkholderia* spp., as the patient ages and disease progresses.<sup>93-94</sup> By early adulthood, many patients have established *Pseudomonas aeruginosa* infections that become progressively resistant to antibiotic therapy.<sup>95-96</sup> Colonization by *Burkholderia cepacia* occurs as the disease advances and coincides with increased morbidity and mortality.<sup>97</sup> Culture-independent methods have also recently identified both aerobic and anaerobic species, such as *Prevotella* spp., *Veillonella* spp., *Clostridium* spp., *Mycobacterium* spp. and *Streptococcus* spp., which had not previously been associated with CF infection and appeared to be harbingers of disease progression.<sup>17, 98-100</sup>

Advances in culture-independent sequencing methods have provided the means to view human microbiota and their associations with health more comprehensively.<sup>9, 15, 101</sup> In the case of CF lung infections<sup>18, 81, 98, 100</sup>, results from these techniques have offered new insights into the complexity of microbial community composition<sup>18</sup> and have begun to challenge conventional wisdom regarding underlying causes of infection, characteristics of disease progression and the efficacy of typical treatment regimens.<sup>93, 95, 102-103</sup> Studies, both in CF and in other environments, have posited the notion that microbial community composition could be a significant determinant of pathogen acquisition.<sup>104-107</sup> Furthermore, a variety of other factors, such as antibiotic treatments, host characteristics and environment, appear to affect pathogen colonization and disease course.<sup>93, 102-103, 108-109</sup>

Our study tracks one adult CF patient over 276 days through different treatments and hospitalizations. We used 454-FLX pyrosequencing and qPCR to characterize the microbiota in 21 sputum samples and to compare treatment response, species abundance and microbial community diversity over time. Our results and previous work emphasize the importance of repeated, longitudinal sampling for studying microbial communities in human subjects<sup>15</sup> where some variation in microbial community composition can occur, even between sequential samples from a single clinically stable patient.<sup>92</sup>

While we recognize that many of our specific findings from our study are not broadly applicable across adult CF patients whose microbiota are distinctive<sup>91-93, 110</sup>, we believe our study results emphasize the utility of regular within-patient sampling and may provide insight into community antibiotic susceptibility, resistance and change beyond this one patient.<sup>80</sup> Furthermore, as sequencing costs continue to decrease, longitudinal sampling and metagenomic analysis is becoming accessible for researchers and clinicians both to inform treatment choices and to guide disease management.<sup>82, 92, 103, 111</sup>

### 3.3 Materials and methods

#### 3.3.1 Patient profile

One 30-yr-old adult male subject, diagnosed with CF at two weeks of age, voluntarily participated in this study of CF microbial community composition and dynamics. He provided informed consent in accordance with an IRB protocol approved by the University of North Carolina at Charlotte. His treatment regimen during the study included oral enzymes for CF-related malabsorption, along with various antibiotics for recurrent respiratory infections since the age of diagnosis. During the sampling period, the subject was also following a prophylactic antibiotic regimen of 500 mg of

azithromycin three times per week and twice-daily nebulizer-administered TOBI® (a solution of 300mg tobramycin and 11.25mg sodium chloride in sterile water), every other month. The subject displayed no other CF-associated complications during the study.

### 3.3.2 Sample collection and preparation

Spontaneously expectorated sputum samples were collected twice weekly for nine months. Twenty-one of the samples were selected for analysis based on the patient's antibiotic regimen, as described in results. They were immediately stored at -20°C to await further preparation. Prior to DNA extraction, samples were pretreated with a 2:1 ratio of sputum to Sputolysin (a solution of 1% 1,4-dithiothreitol, 7% NaOH and 7% H<sub>3</sub>PO<sub>4</sub>), and then incubated at 50°C for 30 min.

### 3.3.3 DNA extraction

A 100uL quantity of sputum was mixed with a 200uL DTT solution and then incubated at 55C for 1 hour to reduce sputum viscosity. Bacterial DNA was then extracted using the IT 1-2-3 VIBE Sample Purification Kit (Idaho Technologies, Salt Lake City, UT). DNA extract was used for quantitative PCR and 16S rRNA gene pyrosequencing.

### 3.3.4 qPCR

The PCR mixture contained 12.5uL Perfecta SYBR Green FastMix Reagent Low ROX (Quanta Biosciences, Gaithersburg, MD), .5uL of 100 pmole/uL of each primer, 5uL of DNA, and 6.5 uL nuclease-free water to a final volume of 25 uL. The primer pair was 27F and 244R with a target size of a 357bp. PCR was performed using the ABI 7500 Fast Real-Time PCR System (Applied Biosystems) with an initial step of 10 min at 95°C, followed by 40 cycles of 15 sec at 95°C and 1 min at 60°C. Melting curves were

determined following the PCR by 1 cycle of 15 sec at 95°C, 1 min at 60°C, 30 sec at 95°C and 15 sec at 60°C.

A standard curve was created using 10-fold dilutions of amplicons generated using the 27F/244R primer pair and an *E. coli* lab strain as the DNA template. DNA copy number per mL of sputum was calculated for each sample based on a standard curve with a  $1 \times 10^6$  fold linear range in CT values.

### 3.3.5 16S PCR for 454-sequencing

The PCR products for 454 tagged sequencing, from extracted previously extracted DNA (see 3.3.3), were prepared with primers and thermalcycling parameters described in Fierer et al.<sup>57</sup> The 454 Life Sciences primer B with a “TC” linker and bacterial 27F primer (5'-GCCTTGCCAGCCCGCTCAGTCAGAGTTTGATCCTGGCTCAG-3') and 454 Life Sciences primer A with a “CA” linker, 12 mer barcode and bacterial primer 338R (5'-GCCTCCCTCGCGCCATCAGNNNNNNNNNNNCATGCTGCCTCCCGTAGGAGT-3'), where the N's represent barcodes used to identify each sample,<sup>112</sup> were used to target the V1-V2 variable regions of the 16S rRNA gene.<sup>113</sup> PCRs were set up with Platinum Taq DNA polymerase (Invitrogen) according to the included protocol with 100ng of bacterial genomic DNA as a template. Each reaction was quantified by PicoGreen on a NanoDrop ND-3300 fluorospectrometer. Samples were pooled in equimolar amounts and concentrated in a vacuum centrifuge before being submitted the Environmental Genomics Core Facility at the University of South Carolina for 454-FLX sequencing.

### 3.3.6 Data and statistical analysis

Close to 270,000 sequences were obtained from 454-FLX pyrosequencing. All sequences were end-trimmed based on the Lucy algorithm at a cut-off of 0.002 corresponding to a quality score of 27.<sup>62</sup> Sequences had to meet the following criteria<sup>114</sup> to be included in the final dataset: (1) an exact match to the 5' primer, (2) Lucy's identified region of poor quality at the 0.002 threshold did not extend beyond the 5' primer, (3) no Ns in the trimmed sequence. The 5' primer (including the barcode) was trimmed from the sequences before analysis. Any sequences that did not meet a length requirement from 180 to 280 bases after trimming were discarded. The 216,677 trimmed, quality-controlled sequences were evaluated for human contamination by using BLASTn<sup>115</sup> searches against the entire bacterial 16S rRNA Ribosomal Database Project (RDP)<sup>58</sup> database. Another 257 sequences did not meet our BLAST e-value threshold of 0.001 and were removed from the dataset.

The remaining 216,420 quality-controlled sequences were assigned to Operational Taxonomic Units (OTUs) at  $\geq 97\%$  sequence similarity using the AbundantOTU2 algorithm<sup>116</sup>, a clustering model that develops consensus sequences that form the basis for OTU clustering. The resulting consensus sequences were aligned using NAST-iEr utility<sup>117</sup> and the resulting alignment was submitted to ChimeraSlayer<sup>117</sup> for detection of chimeric sequences. Any OTU with a corresponding consensus sequence that was determined to be chimeric was removed from the dataset.

We used Fast Unifrac<sup>118</sup> to evaluate whether there were differences between pre-treatment and post-treatment samples based on phylogenetic distances. Using OTU consensus sequences, we mapped the pre- and post-treatment samples and their sequence

frequencies to each OTU and used ClustalX (2.0.12)<sup>119</sup> to align OTU consensus sequences and construct a phylogenetic tree. The Newick formatted tree and mapping file were used in the Unifrac analysis. We conducted the analysis using unweighted parameters and the resulting distance matrix was used to construct hierarchical cluster and Principal Coordinates plots.

We used the SILVA<sup>120</sup> reference tree to develop and visualize a phylogenetic tree of OTU consensus sequences by identifying the best BLAST<sup>115</sup> hit for each OTU consensus sequence to the reference tree and using that branch to stand for the OTU in the tree we constructed. Archeopteryx software (version 0.957 beta)<sup>121</sup> was used to visualize the tree, which we hand-pruned to remove branches not represented by the OTUs in our study. The tree was then “decorated” with the most detailed taxonomic classifications for each OTU consensus sequence that met a minimum classification confidence of at least 80%<sup>60</sup> using the RDP classifier.<sup>59</sup>

Individual sequences were assigned taxonomy using RDP web-based classification software.<sup>59</sup> A minimum 80% threshold confidence score<sup>60</sup> was used to include sequences in each taxonomic group.

The 16S rRNA sequence counts were standardized to the average number of sequences across all samples and logged (base 10) for statistical analysis of **relative** sequence abundance, using the following calculation:

$$\text{LOG}_{10}\left(\left(\frac{\text{Frequency}}{\# \text{ sequences per sample}}\right) * (\text{Average \# sequences per sample} + 1)\right)$$

As an example, consider a dataset in which there was an average of 2,500 sequences per sample assigned to phylum at a 80% confidence level. For one sample within this

dataset, 1,000 sequences were assigned to phylum and, of those, 300 were assigned to the taxon Firmicutes. The transformation would be:

$$\text{LOG}_{10}\left(\left(\frac{300}{1000}\right)*2500+1\right)=2.8756$$

Using this measure, relative abundance corrects for different samples having different total numbers of sequences.

To develop a **weighted** abundance measure, 16S rRNA sequence counts were adjusted, and then logged, to reflect the bacterial counts in each sample from qPCR results using the following calculation:

$$\text{LOG}_{10}\left(\left(\frac{\text{Frequency}}{\# \text{ sequences per sample}}\right)*(\text{qPCR counts per sample}+1)\right)$$

Relative abundance and weighted abundance were then used in qualitative and statistical analyses. Choices of statistical analysis were informed by tests for normal distribution (Shapiro-Wilk W Test) and/or equal variance. In cases where assumptions were not met, Wilcoxon Ranked Sum or Welch's ANOVA were used as appropriate.

Statistical analyses were conducted and figures constructed using the R statistical package (R version 2.7.2 (2008-08-25)), JMP® 8.0 software for Microsoft Windows (SAS Institute) and Microsoft® Office Excel 2003.

## 3.4 Results

### 3.4.1 Experimental design

Our study documents microbial community dynamics of CF lung infection in one adult patient over almost a year of infection exacerbation and treatment (TABLE 3.1) and prophylactic management (see Methods). During the study, the patient was treated with three commonly prescribed antibiotics<sup>88</sup>: Ciprofloxacin, an oral or intravenous broad-

spectrum fluoroquinolone; Sulfamethoxazole, an oral or intravenous narrow-spectrum drug that can be targeted to *Burkholderia cepacia* in CF; and Tobramycin, an inhaled, narrow-spectrum antibiotic, used extensively for *Pseudomonas aeruginosa* infection (FIGURE 3.1).<sup>122-123</sup> Our experimental design was observational, developing over the course of the study with sampling undertaken to provide a range of conditions, including periods of exacerbation, treatment and relative clinical stability. While our sampling scheme is not unique<sup>80, 124</sup>, our longitudinal focus on the changing size and composition of the sputum microbiome, over a large number of samples and range of conditions, is. In addition, we have attempted to overcome the drawbacks of using a single microbial community assessment method by combining quantitative Polymerase Chain Reaction (qPCR) to evaluate changes in overall bacterial abundance and 16S rRNA pyrosequencing gene surveys to characterize microbiome membership associated with variations in bacterial biomass and treatments.

#### 3.4.2 Quantitative PCR bacterial abundance

As a first step in assessing microbial community dynamics, we quantified bacterial abundance in each sample using qPCR (TABLE 3.2). Overall bacterial counts, measured by qPCR, displayed surprising differences between samples, showing a 1.2E+06 dynamic range where the standard deviation was larger than the mean (range = 1.70E+04 – 1.24E+06; mean =  $2.39E+05 \pm 2.90E+05$  s.d). Based on these results, we next examined whether or not acute antibiotic treatment cycles affected these abundance fluctuations.

Because treatment timing and selection cannot be experimentally controlled in a patient, we recognized that the range and overlap of treatments in our study created the potential for confounding. As an attempt to address this issue, we chose to evaluate acute

antibiotic treatment only and assigned each of our samples to one of two treatment groups for our first analysis. Due to the nature of *Pseudomonas aeruginosa* infection in our patient, our patient's inhaled Tobramycin therapy was not included in our definition of acute antibiotic treatment. While Tobramycin is often prescribed to reduce *Pseudomonas aeruginosa* bacterial load in newly established or mild infection, it is primarily used for maintenance therapy in cases of established *Pseudomonas aeruginosa* colonization, such as that in our patient.<sup>123, 125</sup> We therefore defined the acute antibiotic treatment group to include any sample obtained during or immediately after Ciprofloxacin and/or Sulfamethoxazole therapy. All other samples were assigned to the "no acute treatment" group. We then examined the time series based on these assignments (FIGURE 3.2). Based on statistical testing, the two groups demonstrated a significant difference in sample means (FIGURE 3.3: Student's T-test  $P = 0.0090$ ), offering an indication that the sputum microbiome responded to acute antibiotic therapy with a reduction in abundance and, then, quickly recovered after the antibiotics cleared the patient's system.

### 3.4.3 Pyrosequencing

To further characterize the microbial community within each sample, we used 454-FLX pyrosequencing of the 16S rRNA gene to quantify bacterial phylotypes. Sequences were assigned to OTUs, using the AbundantOTU (v2.0) algorithm<sup>116</sup>, at 97% similarity, and OTU frequencies were calculated for each sample. Although the number of distinct OTUs across all samples demonstrated a polymicrobial microbiome, the observed concentration of relative OTU abundance in a small number of OTUs highlights the dominance of a few phylotypes in our patient's lung environment (FIGURE 3.4).

We were interested in understanding the phylogenetic relationships between the OTUs present in the sputum microbiome. Using the Silva reference tree, we constructed a phylogenetic tree of OTU consensus sequences and labeled each branch with the most detailed RDP classification with a confidence score  $\geq 80\%$ . The resulting tree (FIGURE 3.5) showed that, in general, OTUs grouped by phylum or class within phylum and that *Firmicutes* contained the highest number of OTUs, even though the OTUs classified as *Proteobacteria* contained almost 97% of all sequences. *Proteobacteria* also contained the two most abundant OTUs classified to genus, *Pseudomonas* and *Burkholderia*, the most common opportunistic pathogens in the adult CF sputum microbiome.<sup>84, 88</sup> Several OTUs could not be classified to genus, with two of the more abundant OTUs (OTU6 and OTU7) only classified to the domain level.

#### 3.4.4 Microbiome diversity

As a measure of overall microbial diversity in our samples, we computed Shannon-Wiener's Diversity index for each sample (TABLE 3.2), calculated using those OTUs with  $> 5$  sequences across all samples. Our results revealed that samples were not uniformly diverse, with indices ranging between 0.1467 and 1.3579. We then assessed whether there was a relationship between qPCR abundance and microbiome composition. A regression of each sample's diversity index against qPCR abundance showed that the relationship between abundance and diversity ( $R^2 = 0.3424$ ,  $P = 5.324e-3$ ) was positively correlated (FIGURE 3.6A), indicating that conditions which had an effect on overall bacterial abundance, also influenced the diversity of the microbiome. Comparing mean diversity indices between those samples that were associated with antibiotic treatment and those that were not, Welch's ANOVA (unequal variances) confirmed that a primary

impact of antibiotic treatment was a decrease in diversity (FIGURE 3.6B), validating similar results seen in childhood infections exposed to antibiotic treatment.<sup>93</sup>

The antibiotic effects on microbial abundance and diversity in our patient were not unexpected but were noteworthy because Shannon-Wiener's diversity index includes the relative abundance of OTUs, and the two most abundant OTUs dominated the microbiome in every sample. Unifrac<sup>63</sup>, a tool that assesses sample similarity based on phylogenetic distances between sequences, provided a technique that allowed us to assess antibiotic treatment separately from abundance measures. To evaluate the effects on the presence of bacterial phylotypes from a full course of acute antibiotic therapy, we used Fast Unifrac<sup>118</sup> without abundance weighting to calculate a distance matrix that compared the sequence similarity of OTUs from samples taken either directly prior to acute antibiotic treatment or immediately after treatment. The results showed that samples from the two treatment groups clustered together (FIGURE 3.7A), and that all but one pre-treatment sample demonstrated separation in Principal Coordinates Analysis from the post-treatment samples (FIGURE 3.7B).

#### 3.4.5 Microbial community composition

To better understand the types of bacteria that accounted for these diversity shifts, we taxonomically classified each 16S rRNA sequence using RDP.<sup>59</sup> For our analysis, we only used those sequences that had a minimum RDP classification confidence of 80% at each taxonomic level.<sup>81</sup> Our results revealed that *Pseudomonas* and *Burkholderia* overwhelmed other community members as a proportion of the overall sputum microbiome (FIGURE 3.8), comprising over 93% of all sequences for 19 of our 21 samples. *Pseudomonas*, by itself, dominated all other types of bacteria in every sample,

comprising over 90% of total sequences in 13 of the 21 samples obtained. Although other studies that used a variety of methods have shown that *Pseudomonas* is often the major player in the adult CF microbiome<sup>81, 84, 88, 96</sup>, we were surprised by the degree to which relative *Pseudomonas* abundance defined the sputum microbiome in our patient. Nevertheless, hierarchical clustering of samples based on RDP classifications shows that microbes other than *Pseudomonas* drove the similarities and differences that defined how our samples grouped (FIGURE 3.9).

We next examined the sampling time series more closely, visualizing changes in the *Pseudomonas* and *Burkholderia* relative abundance compared to qPCR bacterial counts (FIGURE 3.10A-B). Differences between the two genera in the pattern of change, for both degree and direction, were clearly apparent. Standard measures of variation (TABLE 3.3) revealed that *Pseudomonas* relative abundance differed little from one sample to another (CV = 0.14) compared either to changes in *Burkholderia* (CV = 1.15) or in other bacteria as a group (CV = 1.37). One interpretation of these statistics is that *Pseudomonas* has much higher resistance to antibiotic therapy than does *Burkholderia*, which was supported by our sequencing data but contradicts evidence from recent studies.<sup>126-127</sup> To argue for *Pseudomonas* relative stability also seemed counterintuitive in the context of qPCR bacterial counts that varied by orders of magnitude.

Artifacts of relative abundance measurement could also account for the observed differences in variation.<sup>62</sup> Relative abundance, by definition, requires that when one variable changes, another must change in the opposite direction, which may mask actual variation in a dominant variable, such as we observed with *Pseudomonas* in our patient's sputum microbiome. Indeed, when we conducted regression analysis of *Pseudomonas*

relative abundance against *Burkholderia*, the significantly strong negative correlation ( $R^2 = 0.6037$ ,  $P = 3.43E-05$ ) highlighted a reciprocal relationship between *Pseudomonas* and *Burkholderia* relative abundance which supported this assessment (FIGURE 3.11A).

Defining the relationship between our results based on relative abundance from sequencing versus the findings from absolute abundance from qPCR appeared to be the next logical analytical step. By multiplying the proportion of sequences assigned to each genus in a sample by that sample's qPCR counts, we calculated a weighted abundance measure designed to adjust the relative sequence frequency to reflect qPCR abundance. Comparing summary statistics for relative abundance and weighted abundance (TABLE 3.3), we observed that the large variation between samples in qPCR results was not accurately reflected by our sequencing results alone. The coefficient of variation (CV) for *Pseudomonas* weighted abundance was much higher than that for relative abundance, indicating that *Pseudomonas* might not be as unresponsive to antibiotic treatment as relative abundance results would have suggested. In fact, those samples which were associated with acute antibiotic treatment showed significant differences in *Pseudomonas* weighted abundance as opposed to samples not linked to therapy (FIGURE 3.12A). *Pseudomonas* and *Burkholderia* also demonstrated similar patterns of weighted abundance in regression analysis, opposite from the relationship identified using relative abundance (FIGURE 3.11B).

Based on weighted abundance, *Burkholderia* still remained more responsive to changing conditions than did *Pseudomonas* (TABLE 3.3), showing wide swings in weighted abundance levels with acute antibiotic treatment (FIGURE 3.12B). These findings indicated the temporary effectiveness of acute antibiotic therapy in reducing

*Burkholderia* abundance; however, abundance quickly rebounded, highlighting resilience as a key characteristic of *Burkholderia* infection in our patient.

Despite the relative dominance of *Pseudomonas* and *Burkholderia*, other microbial community members were not insignificant components of the microbiome (TABLE 3.4). Based on taxonomic classification of individual sequences, twenty-nine different genera were classified at an 80% confidence level; however, only three, *Pseudomonas*, *Burkholderia* and *Streptococcus*, were present at all sampling time points. Increases in relative abundance of *Streptococcus* during exacerbation events (FIGURE 3.13A) are noteworthy and support previously reported observations linking increases in *Streptococcus* spp. to infection exacerbation events.<sup>111</sup> Several other minor microbiome members documented in previous pyrosequencing studies exhibited the same pattern of response to acute antibiotics: *Prevotella*, *Veillonella*, *Bacteroides*, *Lactobacillus* and *Rothia*.<sup>81-82</sup> These observations indicate that minor players in the microbiome may have important ramifications for clinical management in CF patients.

The remainder of the microbiome contained additional previously documented bacteria<sup>81-82, 90-91, 128</sup>, including facultative and obligate anaerobes characteristic of the hypoxic environment thought to be created when *Pseudomonas* abundance and biomass is high (TABLE 3.4).<sup>90</sup> Four specific genera, *Moryella*, *Pasteurella*, *Granulicatella* and *Gemella*, displayed similar patterns of abundance increases in the absence of acute antibiotic therapy (FIGURE 3.13B). A few low prevalence genera, were more commonly associated with other microbiota types in the literature (e.g. *Turicibacter* [oral, fecal], *Bradyrhizobium* [plants], *Sneathia* [vaginal]), but, nevertheless, classified with high confidence in our sputum samples. As a group, both relative and weighted abundance for

the less prevalent bacteria varied tremendously over the course of the study (TABLE 3.3) and showed susceptibility to acute antibiotic therapy (FIGURE 3.12C).

### 3.5 Discussion

Our study is one of many that have described the complex, polymicrobial environment that characterizes the adult CF lung; however, our research is among the first to document CF microbiome dynamism using 16S rRNA to examine microbial composition and the effects of antibiotic therapy in one patient over an extended time period.<sup>19, 111</sup> Sequence analysis techniques allowed us to identify temporal changes in CF microbiome composition and confirmed the presence of newly identified community members<sup>81-82</sup> that may have clinical consequences.<sup>102-103, 111</sup> Our choice to combine 16S rRNA pyrosequencing with qPCR quantification unexpectedly illustrated analytical inadequacies inherent in relative abundance measurement. Correcting this potential limitation of pyrosequencing results becomes an important priority as more researchers and clinicians adopt culture-independent approaches required to fully understand polymicrobial CF infection.

For our patient, the dominance of *Pseudomonas* was the most obvious feature of the microbiota and supported existing evidence that an established *Pseudomonas* presence in the CF lung defines the character of infection from that point forward. Several studies have identified genetic changes<sup>88, 129-130</sup> and increased mutation propensity<sup>124, 131-132</sup> over the course of CF infection that map a progression of decreased virulence and increased resistance as the bacteria adapt to their environment to evade host immune responses.<sup>124</sup> Paradoxically, the unwavering abundance of *Pseudomonas* observed from our pyrosequencing results, which indicated virtually complete antibiotic resistance in our

patient, was called into question when we examined abundance changes in the context of qPCR bacterial counts.

The survival benefits for *Pseudomonas* of genetic adaptation are clear. In addition, mechanisms, such as biofilm formation, that support antibiotic resistance<sup>90, 133-135</sup> may also encourage microbial diversity and proliferation rather than crowding out other species. The wide variety of taxa identified in our samples echoes the findings from previous research that *Pseudomonas*-dominant CF infections support a hypoxic environment where anaerobic species can flourish.<sup>81, 136-138</sup> What remains unclear is how secondary players, both major and minor, affect disease progression and exacerbation severity and frequency.

*Burkholderia*, the second most abundant genus in our patient's microbiota, has long been recognized as a common source of cross-infection in adult CF cases.<sup>139</sup> It is notable for its virulence and resistance to therapy<sup>140-142</sup>, further documented by the resilience we observed over the course of our study. The combination of *Pseudomonas* and *Burkholderia* has also been associated with poor clinical outcomes, and their potential interactions have been hypothesized to promote colonization by additional infectious agents.<sup>138, 143</sup>

The observed increase in *Streptococcus* in our samples during exacerbation events provides a potential example of this phenomenon. Sibley et al. documented a similar relationship between *Streptococcus* spp. colonization and pulmonary exacerbations in one patient that was later confirmed in several adult CF patients admitted to the hospital for treatment.<sup>111</sup> Culture-based diagnostic techniques and TR-FLP were not always effective in identifying *Streptococcus*, leading to recommendations of specialized culture and

culture-independent approaches to improve diagnosis.<sup>111</sup> Our analysis also identified other less abundant microbiome members, such as *Prevotella*<sup>144</sup> and *Veillonella*, that appeared to cycle with exacerbation and treatment. The clinical relevance of these observations is uncertain. While culture-independent studies are beginning to characterize the complexity and dynamics of the CF infection biosystem, the polymicrobial mechanisms by which chronic colonization produces clinically detectable infection and pulmonary damage are only beginning to be appreciated.<sup>98, 145-146</sup>

Like many long-term chronic diseases, CF presents significant challenges for clinical management.<sup>88</sup> Those challenges are intensified by a complex and ever-changing CF microbiome, in which the interrelationships between host, therapeutic and microbial mechanisms that drive community dynamics and clinical outcomes are not well understood. Although time series snapshots, such as ours, can provide useful documentation of CF community members and dynamics, a more comprehensive vision is needed.<sup>98</sup> Modeling approaches that consolidate information from multiple experimental and analytical techniques, such as the metabolic network analysis of *Pseudomonas aeruginosa* pathogenicity and adaptation developed by Oberhardt et al.<sup>147</sup>, may ultimately provide an excellent option for resolving the complexity inherent in CF infection, leading to a more effective application of research insights to clinical management and treatment of CF patients.

TABLE 3.1: Clinical observation and medication stage. Exacerbation column indicates which sampling dates occur during clinical signs of infection. For the columns indicating specific medications: Pre = 1-2 days immediately prior to treatment, Post = 1-2 days immediately after treatment was completed, None = that medication was not part of therapy, Yes = treatment was underway.

<b>Sample Date</b>	<b>Exacer-bation</b>	<b>Tobra-mycin</b>	<b>Cipro-floxacin</b>	<b>Sulfame-thoxazole</b>
28Feb	No	Post	None	None
10Mar	Yes	None	Pre	Pre
11Mar	Yes	None	Pre	Pre
25Mar	No	Pre	Yes	Yes
10Apr	No	Yes	Post	Post
11Apr	No	Yes	Post	Post
28Apr	No	Post	None	None
25May	Yes	Pre	Pre	Pre
26May	Yes	Pre	Pre	Pre
27Jun	No	Post	Post	Post
28Jun	No	Post	Post	Post
25Jul	No	Pre	None	None
26Jul	No	Pre	None	None
12Aug	No	Yes	None	None
17Aug	No	Post	None	None
14Sep	No	None	None	None
15Oct	No	None	None	None
5Nov	Yes	None	Pre	None
16Nov	Yes	None	Yes	Pre
21Nov	No	None	Post	Yes
1Dec	No	None	None	Post

TABLE 3.2: Total bacterial qPCR results compared to diversity. Overall bacterial counts from qPCR of sputum microbiome. Standardized mean difference =  $\mu_k - \mu / \sigma$  and provides a relative measure of dispersion for each sample. Shannon's diversity index indicates the genus level diversity calculated from pyrosequencing results.

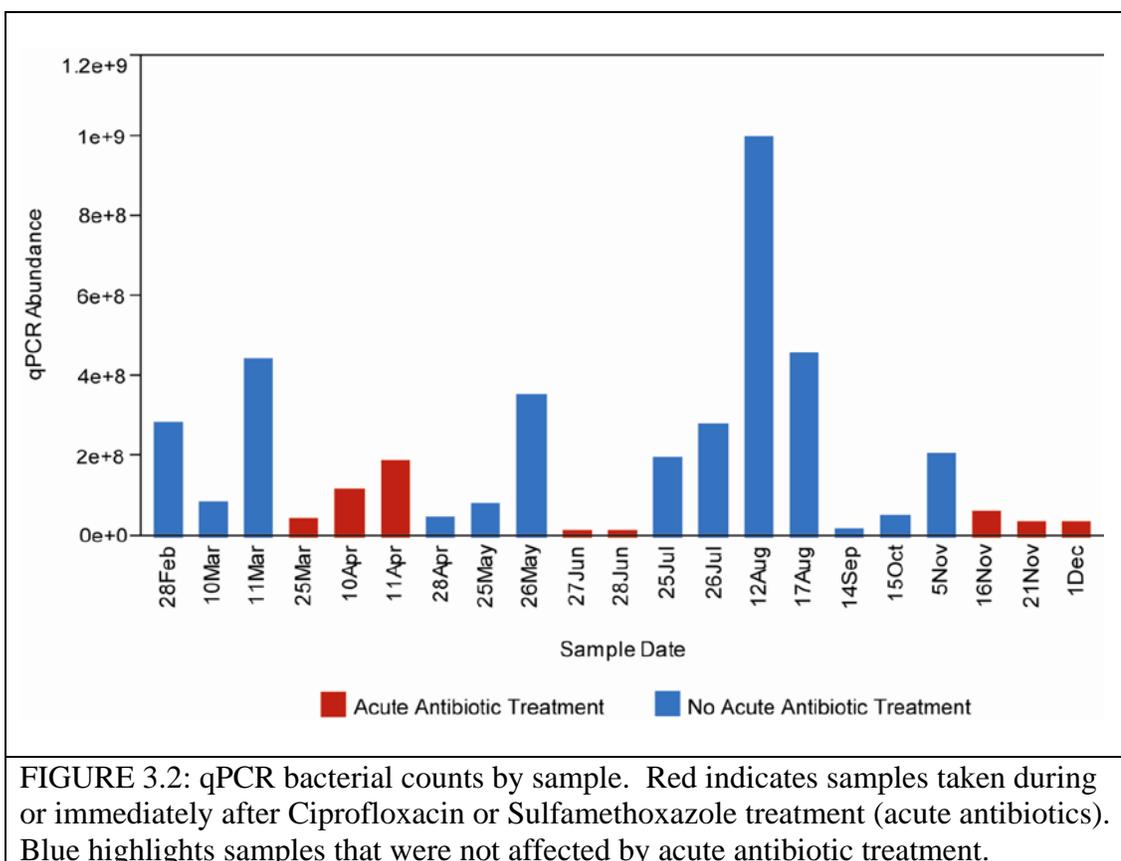
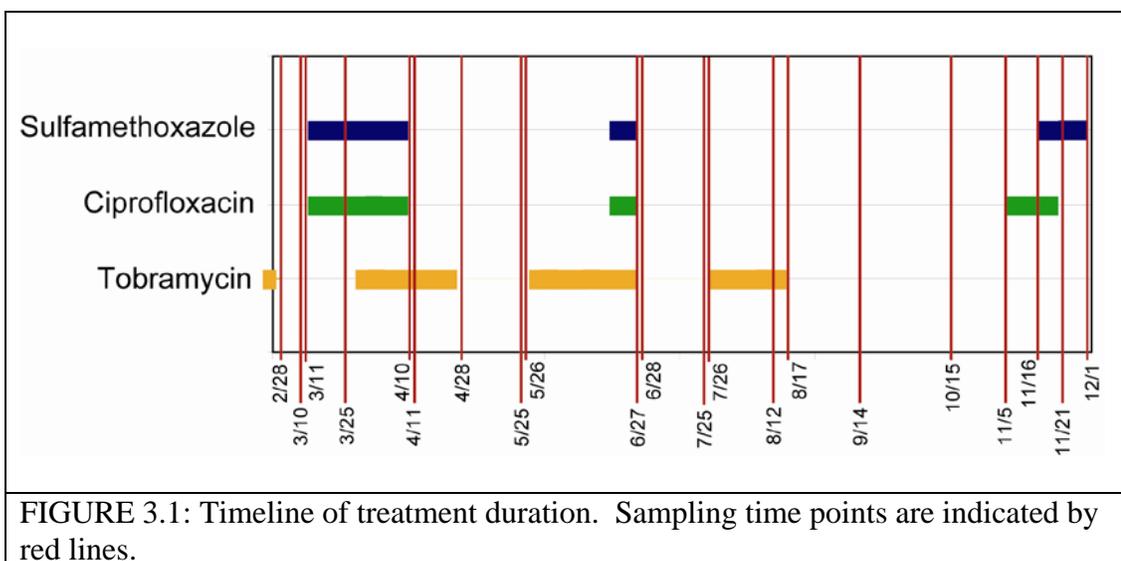
<b>Sample Date</b>	<b>qPCR Abundance</b>	<b>Standardized Mean Difference</b>	<b>Shannon's Diversity Index</b>
28Feb	2.84E+08	3.98E-01	0.3552
10Mar	8.69E+07	-4.53E-01	0.6849
11Mar	4.43E+08	1.08E+00	0.8788
25Mar	4.36E+07	-6.40E-01	0.4268
10Apr	1.20E+08	-3.12E-01	0.2575
11Apr	1.88E+08	-1.51E-02	0.5767
28Apr	4.89E+07	-6.17E-01	0.7016
25May	8.06E+07	-4.80E-01	0.5133
26May	3.54E+08	7.01E-01	1.3579
27Jun	1.53E+07	-7.62E-01	0.3084
28Jun	1.39E+07	-7.68E-01	0.2199
25Jul	1.98E+08	2.59E-02	0.1765
26Jul	2.82E+08	3.90E-01	0.6917
12Aug	9.98E+08	3.48E+00	0.7344
17Aug	4.58E+08	1.15E+00	0.7910
14Sep	1.92E+07	-7.45E-01	0.1969
15Oct	5.28E+07	-6.00E-01	0.5000
5Nov	2.07E+08	6.47E-02	0.4021
16Nov	6.15E+07	-5.62E-01	0.1467
21Nov	3.81E+07	-6.64E-01	0.3915
1Dec	3.68E+07	-6.69E-01	0.2985
<b>Mean</b>	<b>1.92E+08</b>	<b>Standard Deviation</b>	<b>2.32E+08</b>

TABLE 3.3: Comparison of Pseudomonas, Burkholderia and other bacteria relative abundance from pyrosequencing to abundance weighted by qPCR overall bacterial counts. CV = coefficient of variation. Relative abundance uses the number of sequences assigned to the genus or group for each sample based on RDP classification at 80% confidence to calculate statistics. qPCR weighted abundance uses the proportion of sequences represented by each genus or group for each sample applied to the total bacterial counts from qPCR of that sample to estimate qPCR counts for the genus or group. Overall mean, s.d. and CV are calculated from sample results.

Genus	Relative Abundance			qPCR Weighted Abundance		
	Mean	s.d.	CV	Mean	s.d.	CV
Pseudomonas	8027	1095	0.14	2.E+08	2.E+08	1.18
Burkholderia	867	993	1.15	3.E+07	5.E+07	1.75
Other bacteria	329	450	1.37	9.E+06	2.E+07	2.17

TABLE 3.4: Comparison of average LOG10 relative and weighted abundance for genera (> 5 sequences). Relative abundance uses the number of sequences assigned at the genus level for each sample based on RDP classification at 80% confidence to calculate mean. qPCR weighted abundance uses the proportion of sequences represented by each genus for each sample applied to the total bacterial counts from qPCR of that sample to estimate qPCR counts for the genus and calculates the mean from the results.

<b>Genus</b>	<b>Mean Log10 Relative Abundance</b>	<b>Mean Log10 Weighted Abundance</b>
Pseudomonas	5.23	8.18
Burkholderia	4.26	7.48
Lactobacillus	3.53	6.69
Streptococcus	3.10	6.09
Prevotella	2.91	6.16
Veillonella	2.51	5.74
Bacteroides	2.42	5.29
Pelomonas	2.27	5.23
Rothia	2.16	5.44
Turicibacter	2.01	4.59
Bradyrhizobium	1.90	4.79
Moryella	1.70	4.78
Gemella	1.64	4.80
Granulicatella	1.61	4.85
Sneathia	1.57	4.13
Peptoniphilus	1.55	3.90
Oscillibacter	1.46	4.92
Corynebacterineae	1.37	4.32
Parabacteroides	1.32	4.21
Capnocytophaga	1.16	4.36
Oribacterium	1.06	3.41
Blautia	1.02	3.69
Fusobacterium	1.01	3.34



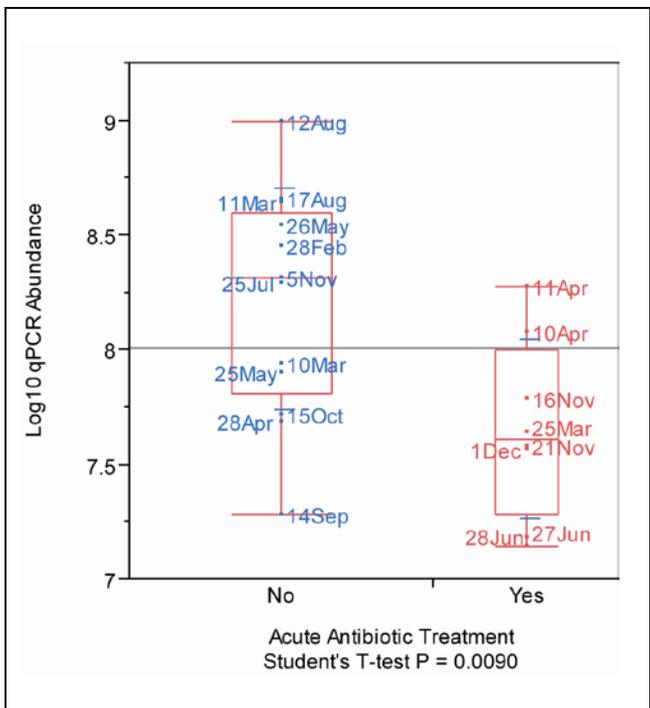


FIGURE 3.3: Test of difference in means between samples in proximity to acute antibiotic treatment and samples not affected.

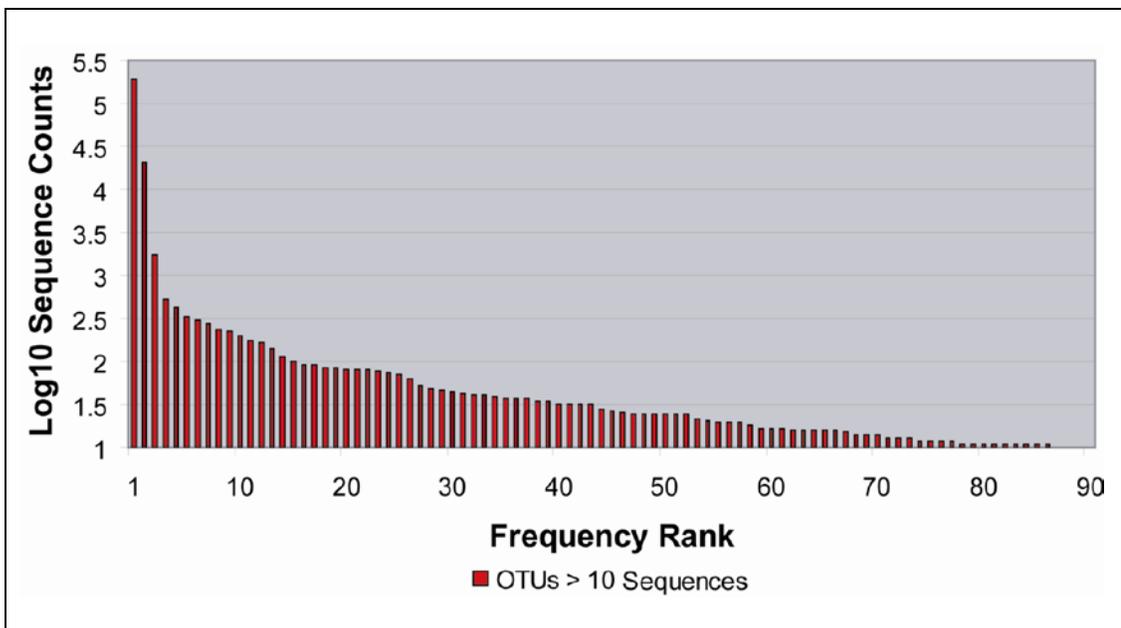
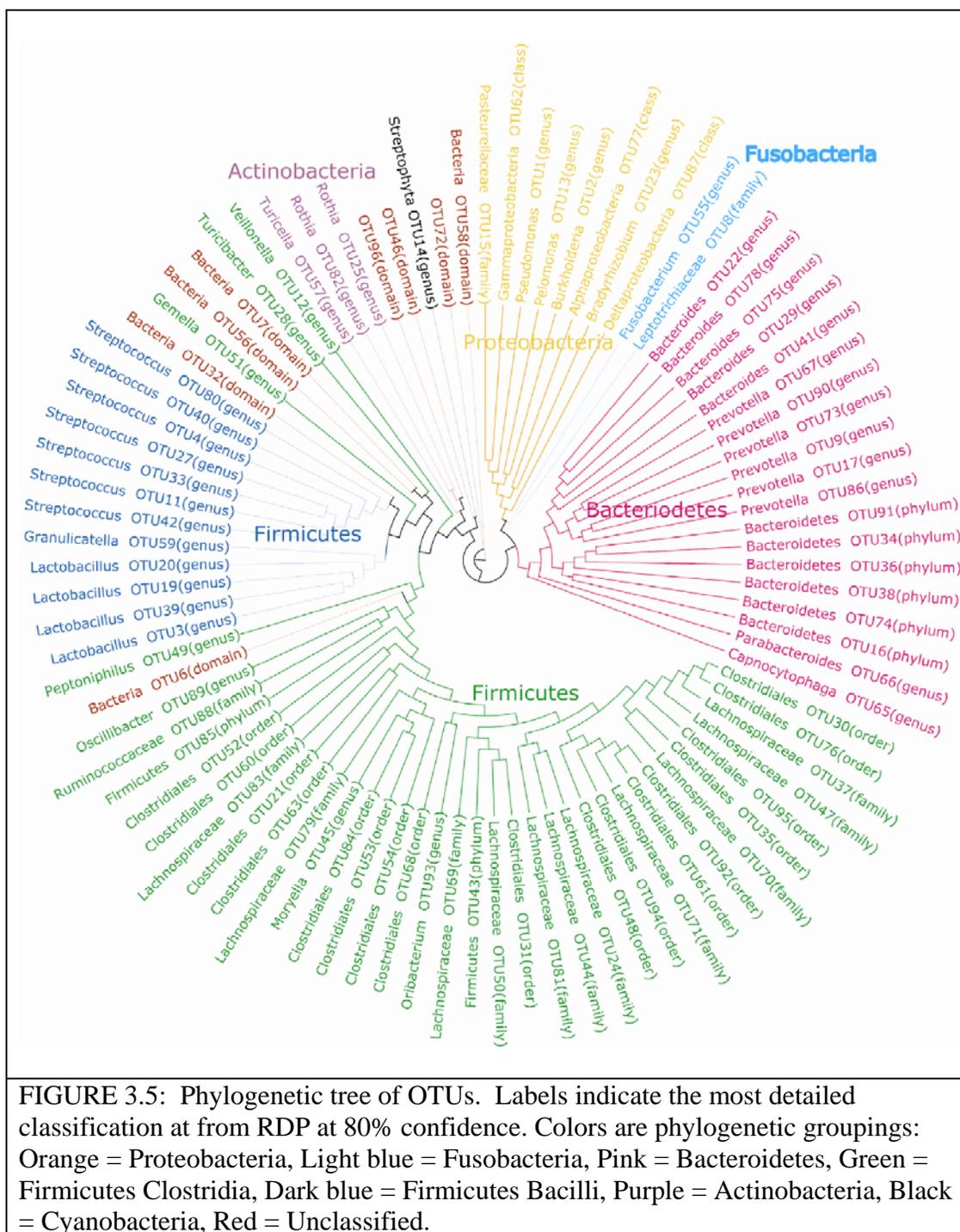
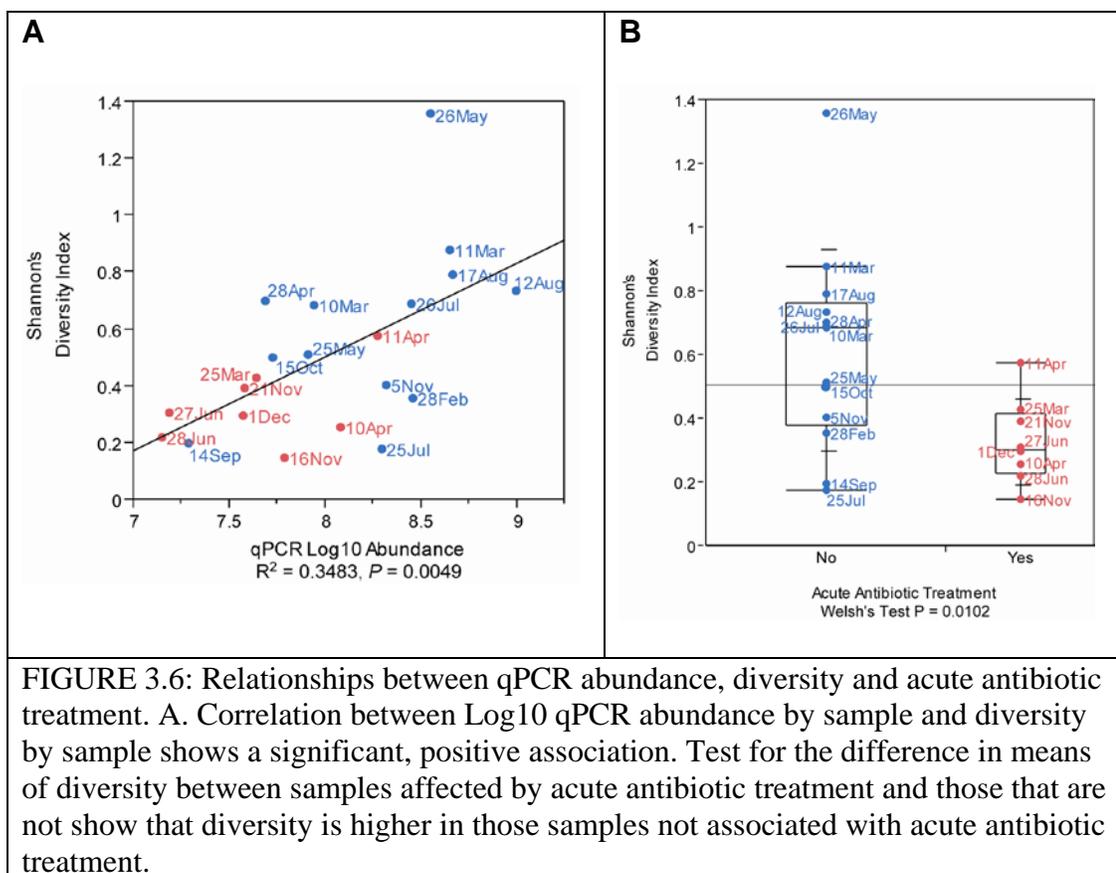
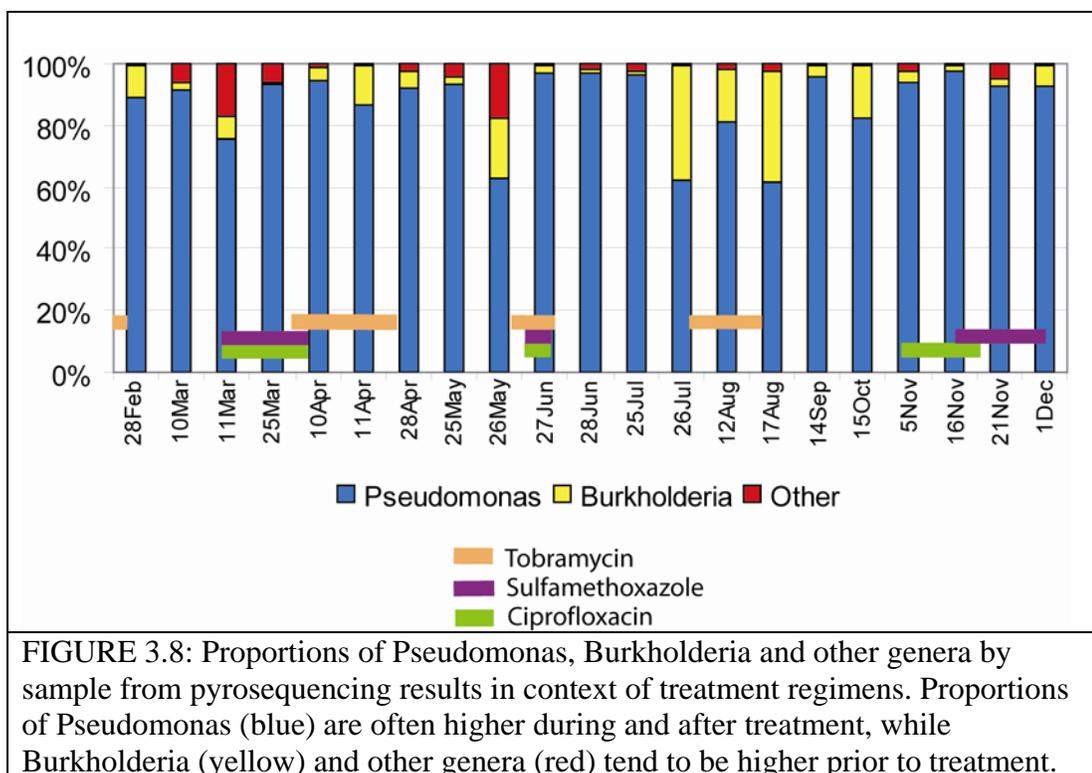
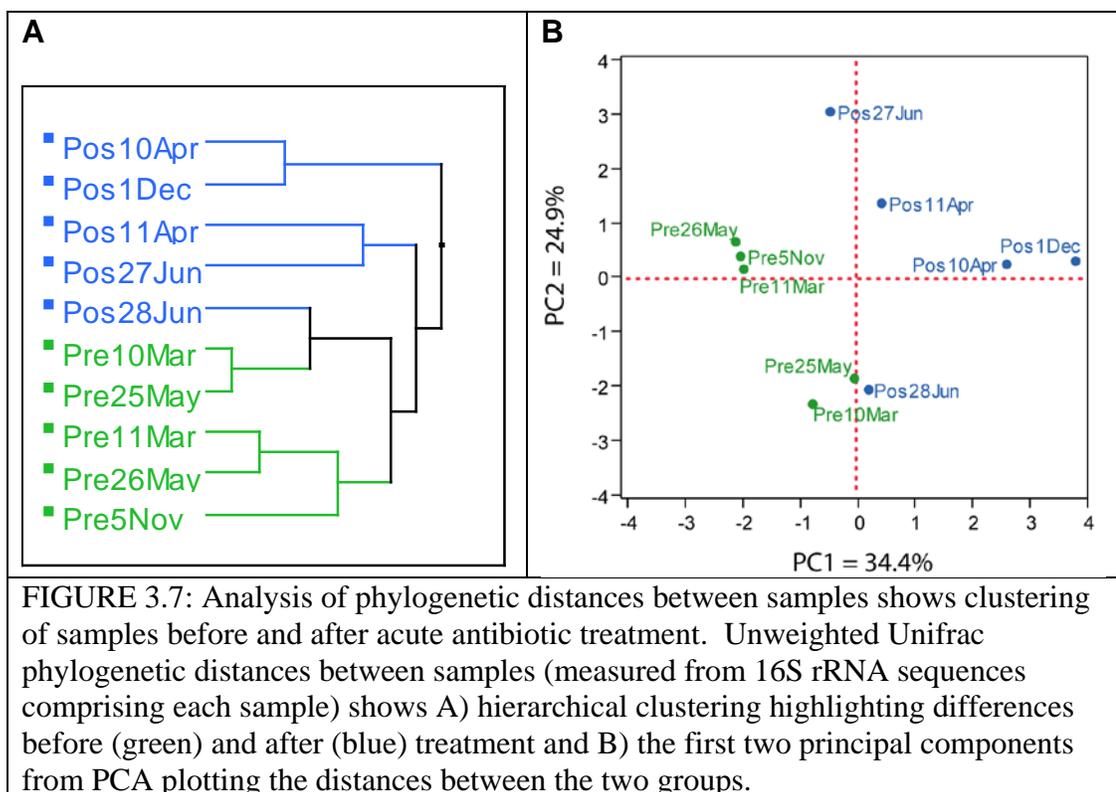


FIGURE 3.4: Log<sub>10</sub> total sequences by OTU ranked in descending order. Sequence abundance by OTU from pyrosequencing illustrates the dominance of the most abundant OTUs.







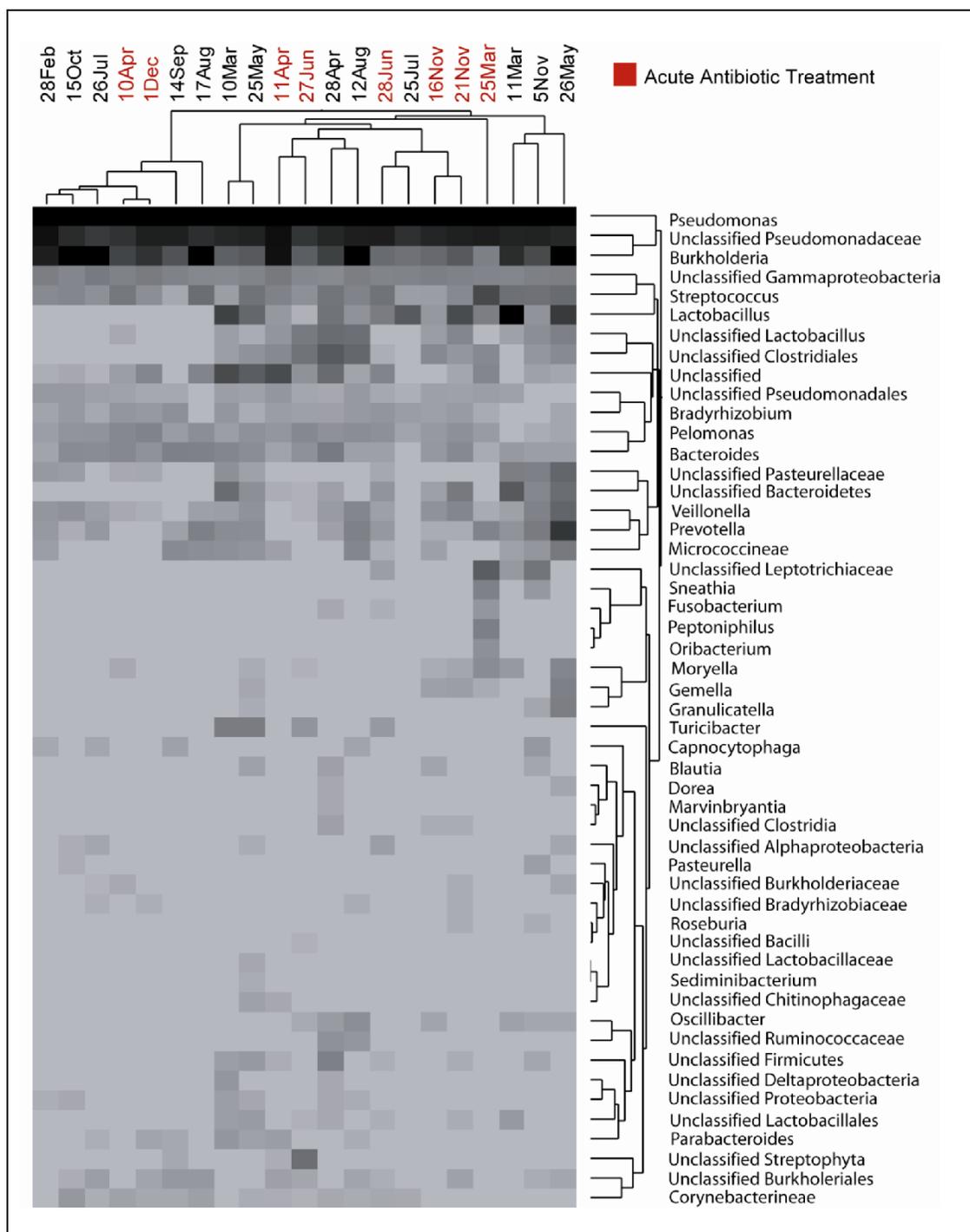
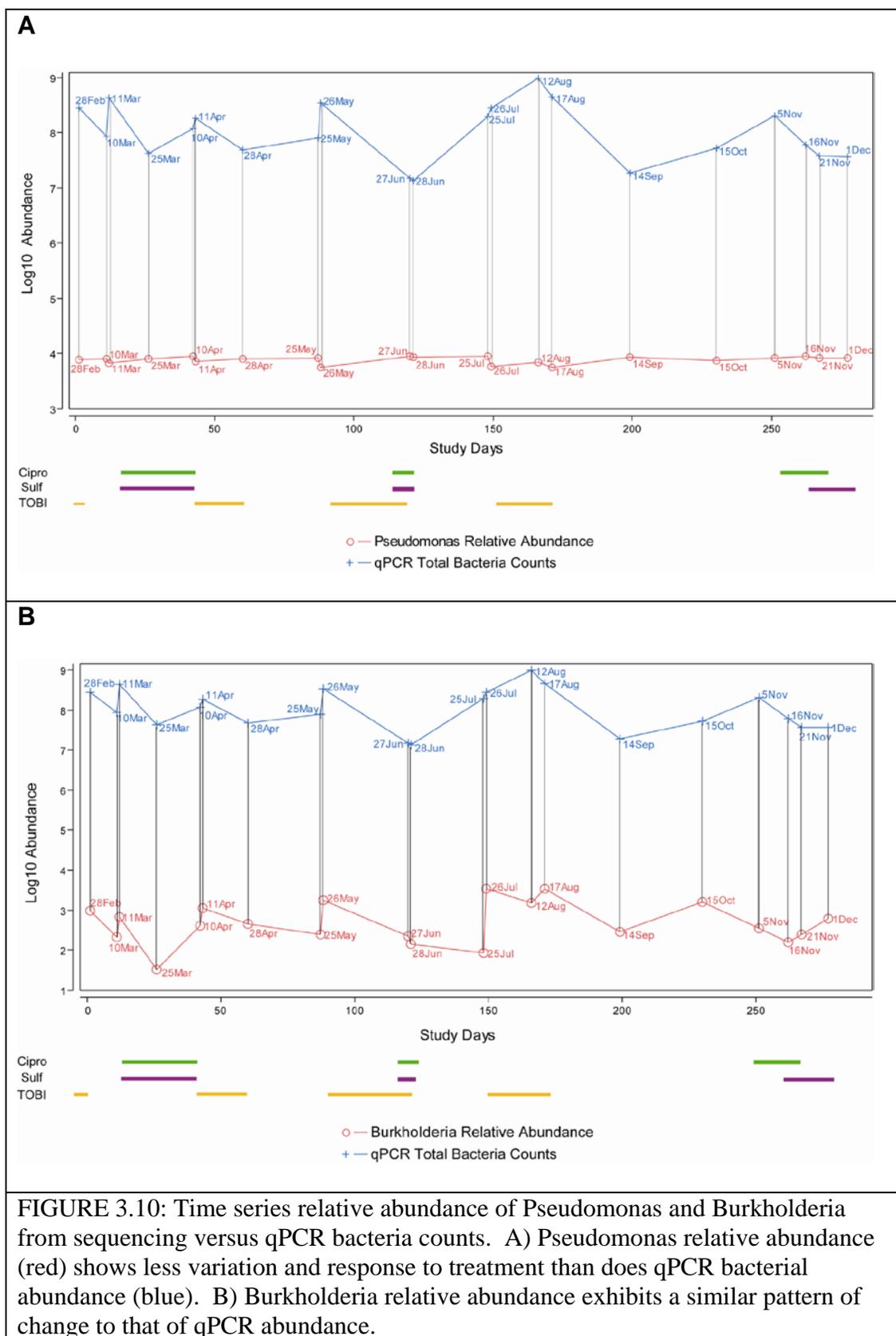


FIGURE 3.9: Two-way clustering of samples based on numbers of sequences grouped by most detailed taxonomic classification at 80% confidence. Samples in red are those during or directly after acute antibiotic treatment.



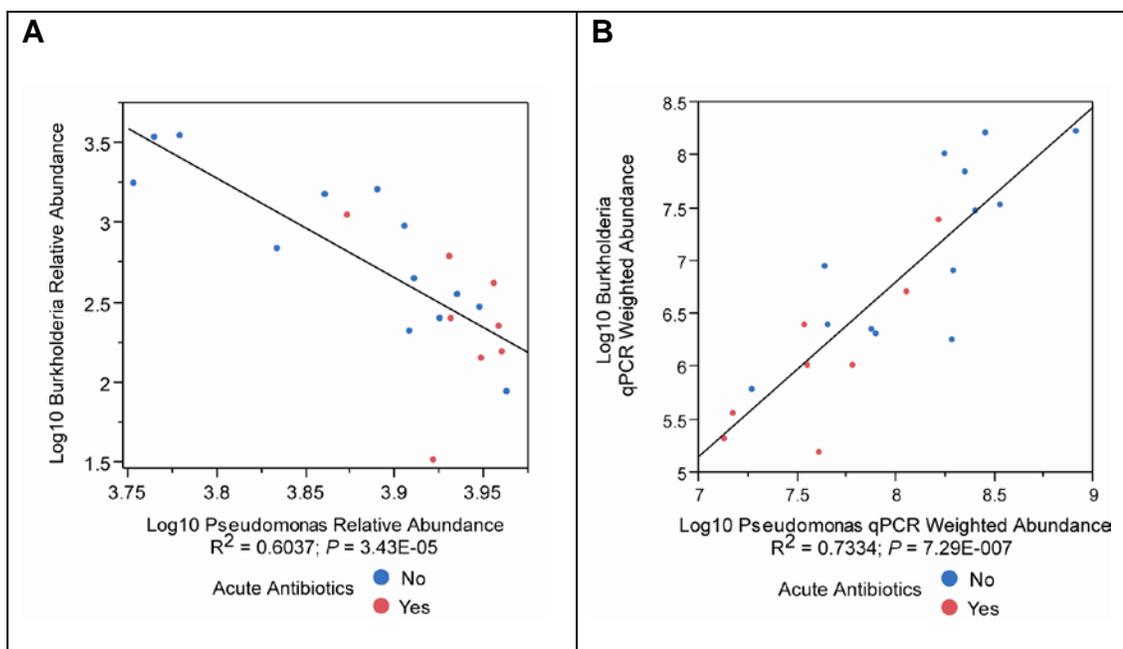


FIGURE 3.11: Associations between *Pseudomonas* and *Burkholderia* abundance. A) Correlation between Log<sub>10</sub> relative abundance of *Pseudomonas* and *Burkholderia* shows a strong inverse relationship. B) When abundance is weighted by qPCR bacterial counts for each sample, an entirely different result is apparent with *Pseudomonas* and *Burkholderia* showing a similar response to conditions at each sampling period. Sample points in red indicate those samples obtained during or immediately after acute antibiotic treatment, and samples in blue are not associated with acute antibiotics.

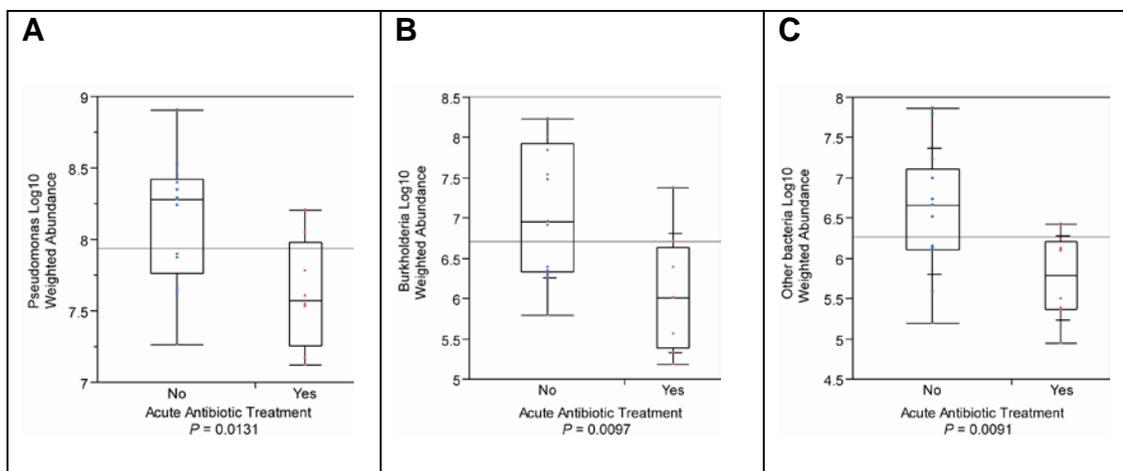
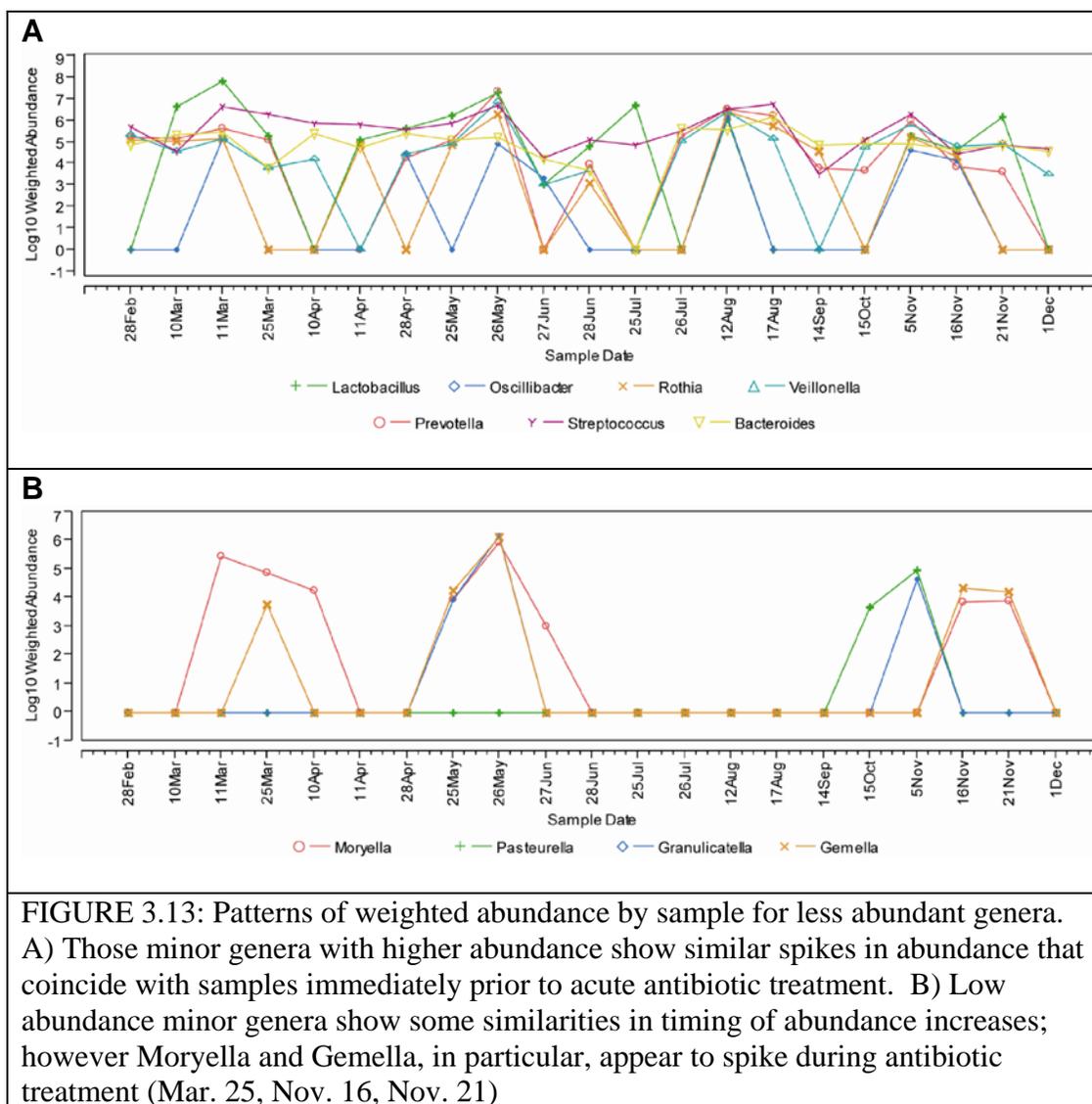


FIGURE 3.12: Test of mean difference in weighted abundance response of bacterial types to acute antibiotic treatment. A) *Pseudomonas*, B) *Burkholderia*, C) all other genera as a group. All differences are significant but *Burkholderia* and other genera appear to exhibit a stronger response.



## CHAPTER 4: AUTOMATED RIBOSOMAL INTERGENIC SPACER ANALYSIS AS AN EFFECTIVE ALTERNATIVE TO SEQUENCING FOR CHARACTERIZING COMPLEX BACTERIAL COMMUNITIES

### 4.1 Background and significance

Snapshot profiling techniques rely on molecular methods, such as T-RFLP, DGGE, LH-PCR and ARISA, and use characteristics of the 16S rRNA, or other genes that can distinguish between types of bacteria, to estimate microbial diversity in metagenomic samples.<sup>19, 110, 148-151</sup> While these methods have significant cost and speed advantages over metagenomic 16S rRNA sequencing, the trade-off is depth of information.

Although microbial community diversity can be estimated and samples can be compared on a relative basis, these techniques do not provide the analytical richness that is available from sequencing. As a result, taxonomic classification of metagenomic samples remains out of reach for many research studies.

The combination of relatively high sequencing expense and cheap, low resolution snapshot techniques suggests that a robust way to associate these methods with specific taxa<sup>19</sup> could offer a much higher resolution analysis of metagenomic samples at a reasonable cost with results generated within 1-2 days. ARISA (Automated Ribosomal Intergenic Spacer Analysis),<sup>56</sup> which relies on differences in the length of the intergenic region in different bacterial taxonomies to characterize a complex microbial community (FIGURE 4.1), produces a electropherogram, a spectrum where each peak indicates the

presence (and qualitative abundance) of a specific community member, although the taxonomic classification of that member remains unknown.

This study investigates whether results from ARISA combined with 16S rRNA Sanger sequencing<sup>152</sup> on the same metagenomic sample could provide enough information to identify a relationship between sequence-derived taxonomic classifications and ARISA peaks. With an initial investment in sequencing the intergenic regions between the 16S and 23S rRNA genes from a series of metagenomic samples, taxonomic classifications could be associated with the intergenic lengths which could, in turn, be assigned to the corresponding ARISA peak. Similar approaches have been attempted for LH-PCR.<sup>19</sup>

If successful, the assignment of ARISA peaks to taxa in this manner would offer metagenomic investigators a rapid-turnaround, inexpensive characterization of a microbial community that could be used to assess sample quality, to quickly compare samples to see if an experiment worked and to determine whether expending additional resources on sequencing would be worthwhile. For example, in a 30-day intervention experiment with 10 subjects, assessing the effects from the intervention on microbial community composition might require sampling every 3 days. 454-FLX pyrosequencing could be conducted on the total of 100 samples at an approximate cost of \$10,000 and the results might not be available until a month after the experiment was completed. If the intervention did not affect the microbial community until the subjects had been treated for 10 days, sequencing the 30 samples that showed no differences would waste resources. Alternatively, ARISA with taxonomic peak assignments would allow each sample to be assessed for response to treatment within 1-2 days of sampling at a cost of

\$2 per sample. The result would be more efficiently allocated resources and more timely information on experimental effects.

## 4.2 Methods

### 4.2.1 Sample collection and preparation

A stool sample was obtained from a female patient undergoing choline depletion as part of a study being conducted by Dr. Steven Zeisel at The University of North Carolina at Chapel Hill. The sample was frozen at  $-80^{\circ}$  and shipped on dry ice to The University of North Carolina at Charlotte where it was stored at  $-20^{\circ}$ . The sample was thawed on ice and metagenomic DNA was extracted and fragmented using QIAamp DNA Stool Mini-kit. Extracted DNA was then refrozen at  $-20^{\circ}$  to await further preparation.

### 4.2.2 ARISA sample preparation and analysis

ARISA was used to detect microbial signatures in our sample based on taxonomic differences in the intergenic distance between the 16S and 23S ribosomal RNA genes (FIGURE 4.1). Using extracted DNA, the intergenic region between the 16S rRNA and the 23S rRNA genes was PCR amplified for each of two technical replicates using a universal 16S forward primer and a eubacterial 23S reverse primer which was fluorescently tagged. The labeled fragments were then separated by size on an Applied Biosystems 3130 sequencer, resulting in an electropherogram where each signal peak represented a different “operational taxonomic unit”. The electropherogram from each replicate was analyzed using internally-developed JAVA software<sup>153</sup> that establishes a base-pair scale, calls the data peaks from the image, assigns base pair sizes to each peak and provides summaries of the peak call assignments. Peak calls were validated using ABI 3130 analysis software.

#### 4.2.3 Sequencing sample preparation

Extracted DNA was thawed and used to PCR amplify a DNA region from the 16S to 23S rRNA genes with the 967F 16S primer and 125R 23S primer using Platinum Taq polymerase from Invitrogen. The resulting PCR products were inserted into pCR 2.1-TOPO vectors (Invitrogen) and sent to Washington University where they were transfected into *E. coli* K12. *E. coli* colonies were then randomly picked and grown in media. Sequences were extracted by mini-prep and the following 3 primers were used to perform sequencing reactions: M13R CAG GAA ACA GCT ATG ACC (vector primer), T7 promoter TAA TAC GAC TCA CTA TAG GG (vector promoter) and 1406F/TGY ACA CAC CGC CCG T (16S rRNA primer). Resulting fragments were sequenced on Applied Biosystems 3730XL. A file containing 1152 sequence fragments representing 384 potential contigs and their trace files was created and sent to the University of North Carolina at Charlotte for analysis.

#### 4.2.4 Sequence assembly and taxonomy assignment

Sets (3 sequences per vector) of matched sequences were analyzed for sequencing quality. They were then aligned and assembled using CodonCodeAligner software (Copyright © 2009 CodonCode Corporation),<sup>154-156</sup> both with and without end trimming (FIGURE 4.2). Resulting assembled contigs (361) were NAST-aligned<sup>157</sup> and checked for chimeric sequences using Greengenes Bellerophon (Version 3).<sup>158</sup> Chimeras were identified in 39 of the assembled contigs which were removed from the analysis. Trimmed and untrimmed contigs were classified for sequence taxonomy using the RDP (Ribosomal Database Project) classification tool.<sup>59, 159</sup> Because of excessive end-trimming, trimmed sequence classification resulted in a significantly reduced analysis set

which appears to be due to removal of portions of the V6 region in the trimming process. Of the untrimmed sequences, 274 were successfully classified.

#### 4.2.5 Intergenic spacer length identification

To determine the intergenic sequence between the 16S and 23S rRNA genes in each assembled contig, four forward and twelve reverse primer sequences (includes ambiguous nucleotide assignments and reverse complements from the degenerate primers used in PCR) from the ARISA analysis were used as search sequences against assembled contigs. Only exact matches were considered to be a hit. From the positions of the matched primers, intergenic lengths between the primers were calculated for each sequence. JAVA software was developed for the primer searches, intergenic length calculations and assignment of taxonomy to intergenic length. Intergenic lengths were verified visually on sample contig sequences.

#### 4.2.6 Analysis

ARISA peak assignments and contig intergenic length frequencies were normalized by dividing each signal intensity or frequency by the sum of all signal intensities or frequencies for the data set. The data was plotted using IGOR software. Correlations were calculated using the R statistical package (R version 2.7.2 (2008-08-25), Copyright (C) 2008 The R Foundation for Statistical Computing ISBN 3-900051-07-0).

#### 4.3 Results

Two types of metagenomic analysis were applied to the same fecal sample from a female patient undergoing choline depletion, ARISA and 16S rRNA Sanger sequencing. ARISA was first used to examine microbial diversity in the sample. The spectrum on an ARISA electropherogram displays peaks that are associated with specific intergenic

lengths of the PCR products that have been run through the sequencer. A standards ladder comprised of fragments of known intergenic lengths is used to calibrate the ARISA signal and assign base pair lengths to each of the peaks in the ARISA spectrum. The intensity of each peak measures relative abundance. Our results showed that the sample contained two dominant peaks at approximately 690 and 729 in base pair lengths (FIGURE 4.3). Several other peaks were of intermediate intensity. Each of these peaks demonstrated the presence of one or more microbial taxa in the sample.

In the second analysis, metagenomic DNA extracted from the same sample was used to amplify a region of microbial DNA spanning from position 967 in the 16S rRNA gene to position 125 in the 23S rRNA gene. Sanger sequencing was performed using 3 different primers creating overlapping sequences that spanned from the 16S V6 region to the 23S rRNA gene (FIGURE 4.2). For 384 clones, we generated 3 sequences each, one for each primer, for a total of 1152 sequences that were assembled into 384 contigs. These contigs then underwent quality checking and removal of chimeric sequences (TABLE 4.1). The final set of assembled contigs was then used in additional analysis steps.

To determine the intergenic length from the end of the 16S rRNA gene to the beginning of the 23S rRNA gene, primers used to synthesize the intergenic spacers for the ARISA analysis were identified within each assembled contig and intergenic sequence lengths were calculated for each sequence. The frequencies of intergenic lengths from the assembled contigs were not evenly distributed in the sample. Two lengths, at 700 and at 739, were dominant. The resulting intergenic lengths were associated with the RDP classifications identified for each sequence (TABLE 4.2). In

several cases, multiple classifications were linked to the same intergenic length and, in other cases, several different intergenic lengths were assigned to the same RDP classification.

ARISA peak assignments and calculated contig intergenic lengths were then compared to determine if there was any correspondence between the results from each method. To reduce noise in the ARISA signal, only the peaks with highest signal intensity were used in the evaluation (37 peaks in total). Both ARISA signal intensities and calculated intergenic length frequencies were normalized to ease the comparison. A normalized ARISA spectrum and the sequence frequencies at each intergenic length were compared (FIGURE 4.4). While the signal intensities showed differences between the two plots, the signals demonstrated a similar peak pattern with the sequence intergenic lengths shifted to the right of the ARISA plot. This correspondence provided some evidence that associating ARISA signal with taxonomy might be possible.

In examining the ARISA peak calls, there appeared to be some consistency in the size of the signal shift between the ARISA signal and the frequencies of the contig intergenic lengths of approximately 10 base pairs. When intergenic lengths at every position were adjusted by +10 base pairs for the ARISA data and the signal was plotted again (FIGURE 4.5), the shift appeared to be quite consistent, with signals from both sets of data aligning well. To determine how well the patterns aligned, correlations were performed on the data with and without correction. The resulting correlation coefficient on data without correction was -0.01175, while the 10 base pair addition substantially improved the R-value to 0.54910.

#### 4.4 Discussion

ARISA is a blunt instrument because, even though each peak in an ARISA spectrum putatively represents a different taxon in a sample, the taxonomic assignment for each peak is unknown. The low resolution of ARISA and other snapshot analysis techniques has been a limiting factor in their utility for detailed taxonomic analysis. Nevertheless, the cost of ARISA for metagenomic analysis compared to sequencing, approximately \$2 per sample versus approximately \$500 per sample for Sanger sequencing, argue for an evaluation to determine if their output could have value beyond simple diversity assessment.

In this study, we have attempted to associate an ARISA spectrum and Sanger sequencing results from a single sample using intergenic lengths calculated from both methods. The ultimate goal of this exercise is to determine whether or not a library of taxonomic classifications for human gut microbial sequences can be associated with ARISA intergenic spacer lengths. This library would then allow inexpensive characterization of complex metagenomic human samples within a few days of sampling.

To obtain sequences long enough to cover both 16S-23S intergenic distance and the 16S V6 variable region for taxonomic classification, we employed a method that sequenced three overlapping fragments and assembled those fragments. To test our hypothesis that ARISA and sequence could be associated, we only sequenced a small number of DNA fragments. Our process had a yield of 71% (274 net sequences), once low quality sequences and chimeras were removed. We anticipated that the highest intensity ARISA peaks would be an indicator of abundance and that these peaks would

show an association with the calculated intergenic lengths from the assembled contigs. For several of the ARISA peaks, this was the case.

An association between ARISA intergenic spacer length and calculated intergenic sequence length does appear to exist; however, for this experiment, that relationship is not straightforward. Intergenic lengths in sequence contigs demonstrate visually similar patterns to ARISA spectra. This relationship improves substantially when the ARISA spectrum is shifted right by 10 base pairs. Many ARISA peaks with high signal intensity tended to show correspondence to contig intergenic lengths that were most frequent. Several possibilities exist that would explain this shift. The presence of fluorescent tags may have affected the ARISA signal. The condition of reagents could also be a factor in the precision of the instrument. The shift might be a pattern that is intrinsic to the physical properties and chemistry of the method. Without further experimentation, the consistency of this pattern could not be established, and the cause could not be resolved.

The degenerate and redundant RDP classifications across adjacent contig intergenic lengths presented a further problem. RDP classifications that corresponded to our calculated intergenic sequence lengths clearly showed that, in several cases, multiple types of bacteria had the same intergenic length and one classification could have several sets of ribosomal genes with different intergenic sequence lengths (TABLE 4.2). For correspondence between intergenic length and classification to be used to accurately assign spectrum peaks to bacterial types, the assignments could be overly ambiguous. This problem might be partially addressed with a binning strategy and knowledge of which taxonomies are typically found in the samples being used; however, these solutions would require further sampling and sequencing to test whether they would be

effective. Samples with higher levels of diversity than the sample diversity in this experiment could also complicate a binning strategy. With no experimental rationale or precedent in the literature to support shifting the base pair assignments and no reasonable resolution to the problem of redundant and degenerate assignments, we concluded that our experiment did not succeed in accurately assigning ARISA peaks to taxonomies.

#### 4.5 Conclusion

The ability to use inexpensive, speedy ARISA analysis to thoroughly profile human gut metagenomic samples could provide many benefits. It would make metagenomic analysis accessible to moderately-funded studies, would greatly reduce the turnaround from sampling to analysis and would make large sample sizes affordable. Nevertheless, our results indicated that several problems remain unsolved.

TABLE 4.1: Sequencing results from assembled contigs and quality checks.

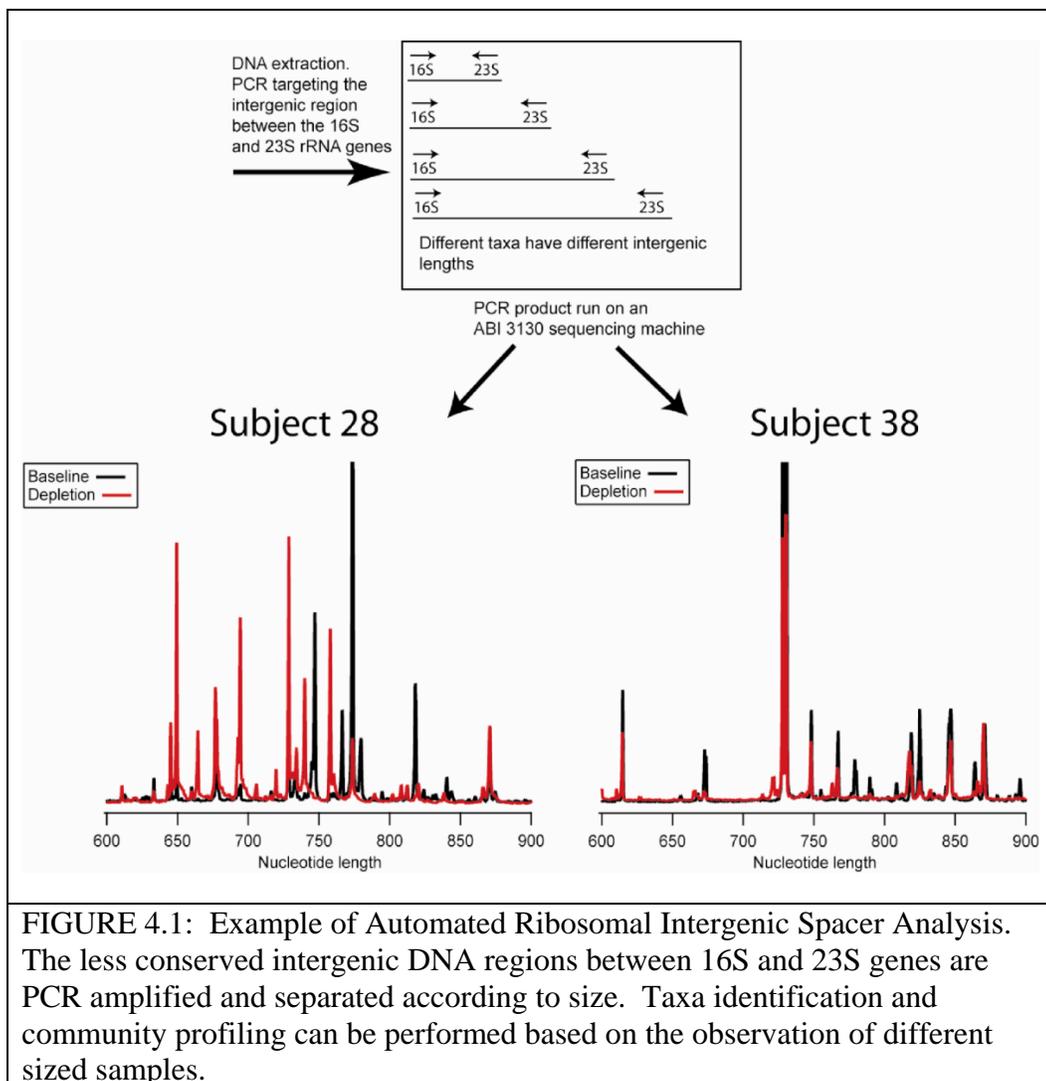
Number of sequence fragments	1152
Number of potential assembled contigs	384
Successful untrimmed assemblies	361
Contigs containing intergenic sequence primer pairs	313
Chimeric sequences	39
Net assembled contigs	274

TABLE 4.2: Intergenic sequence lengths from contig assembly and their corresponding sequence classifications.

Intergenic Length	Freq	Phylum	Class	Order	Family
460, 461	4	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae
487	1	Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae
525, 526	25	Firmicutes	"Clostridia"	Clostridiales	Veillonellaceae
526	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae
551	1	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae
551	1	Proteobacteria	Betaproteobacteria	Hydrogenophilales	Hydrogenophilaceae
571	1	Firmicutes	"Erysipelotrichi"	"Erysipelotrichales"	Erysipelotrichaceae
636	1	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae
638	1	Firmicutes	"Clostridia"	Clostridiales	Incertae Sedis XV
638	1	Firmicutes	"Clostridia"	Thermoanaerobacterales	Incertae Sedis IV
689	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
698	1	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae
698, 699	2	Bacteroidetes	Bacteroidetes	Bacteroidales	Prevotellaceae
699, 700, 701, 702, 703	106	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
700	3	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
700	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae
700	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae
700	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
700	1	Firmicutes	"Clostridia"	Clostridiales	Incertae Sedis XV
700	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae
711	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
725, 728	3	Firmicutes	"Clostridia"	Clostridiales	Veillonellaceae
728	1	Firmicutes	"Clostridia"	Clostridiales	"Lachnospiraceae"
736, 737	3	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
737, 738, 739	9	Bacteroidetes	Bacteroidetes	Bacteroidales	Rikenellaceae
737, 739, 741	4	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae

Table 4.2 (continued)

747	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
767	1	Firmicutes	"Clostridia"	Clostridiales	"Lachnospiraceae"
767	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
772, 773, 774, 775, 776, 777	79	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
774, 775	2	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
774	1	Firmicutes	"Clostridia"	Clostridiales	"Lachnospiraceae"
774	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Prevotellaceae
774	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
798	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
799	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
802	1	Firmicutes	"Clostridia"	Clostridiales	Incertae Sedis XV
811	1	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae
844	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
858	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
876	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
886	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
891	1	Firmicutes	"Clostridia"	Clostridiales	"Ruminococcaceae"
899	1	Firmicutes	"Clostridia"	Clostridiales	"Lachnospiraceae"
1033	1	Firmicutes	"Clostridia"	Clostridiales	"Lachnospiraceae"



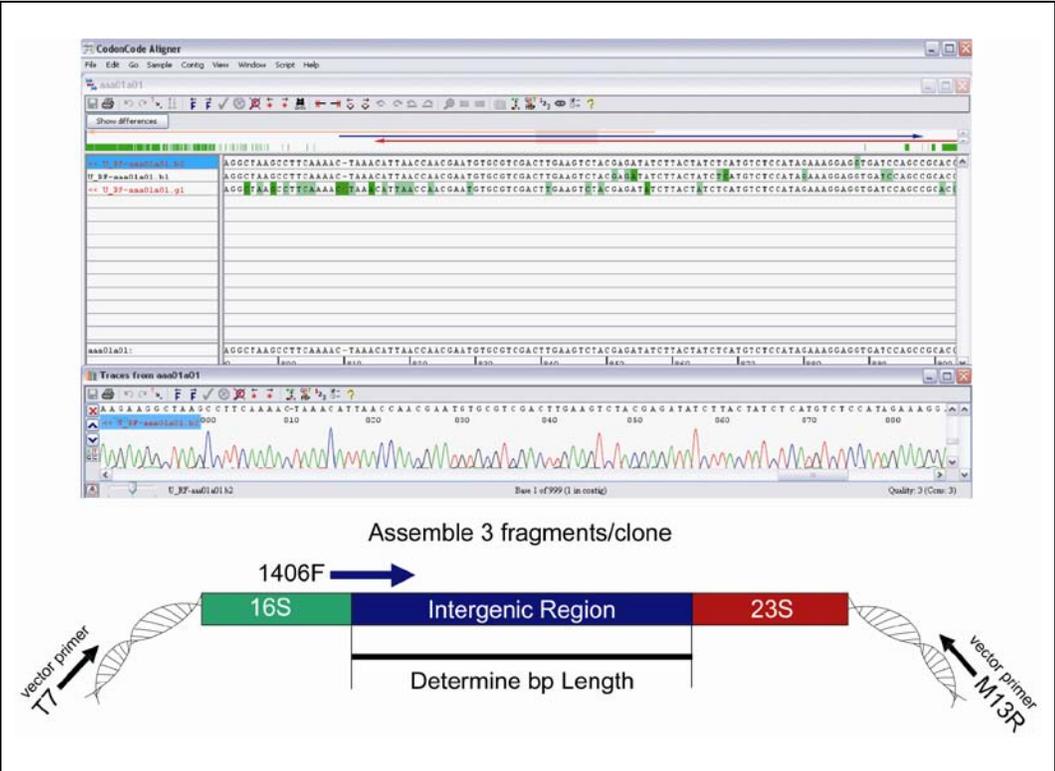


FIGURE 4.2: Example of Sanger sequence contig assembly using CodonCodeAligner.

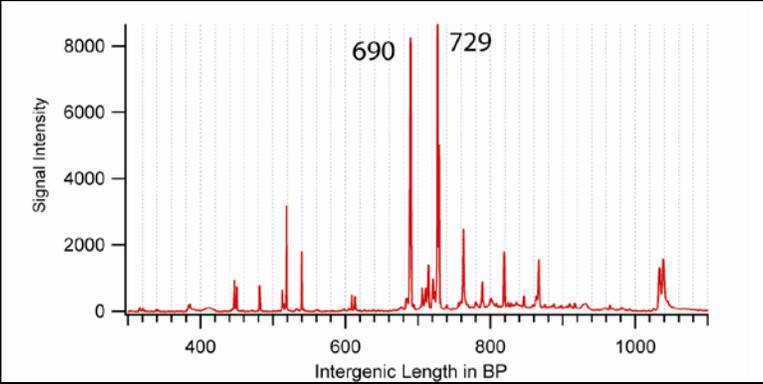


FIGURE 4.3: ARISA spectrum from patient 30 showing two dominant peaks associated with intergenic lengths of 690 and 729.

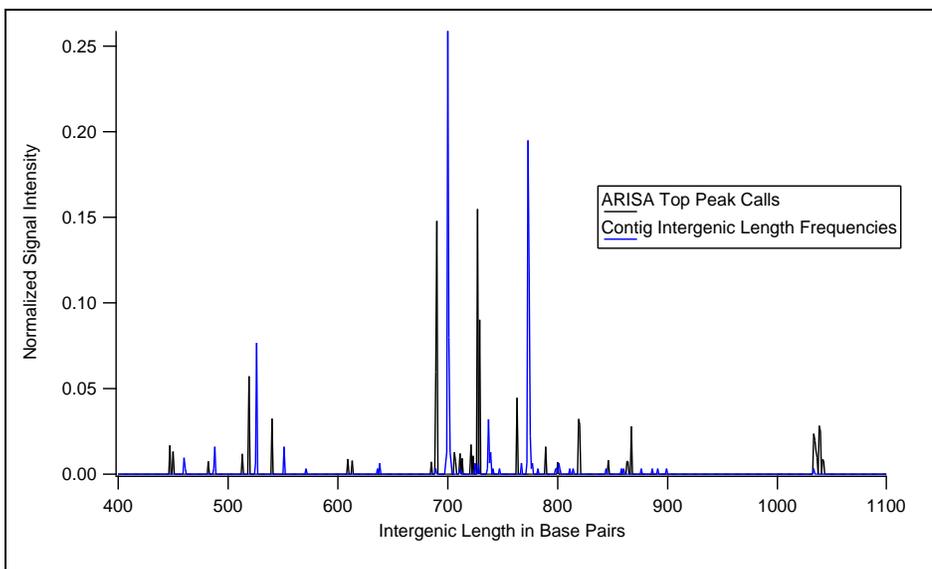


FIGURE 4.4: Comparison of normalized signal from ARISA and intergenic lengths from Sanger sequencing assemblies.

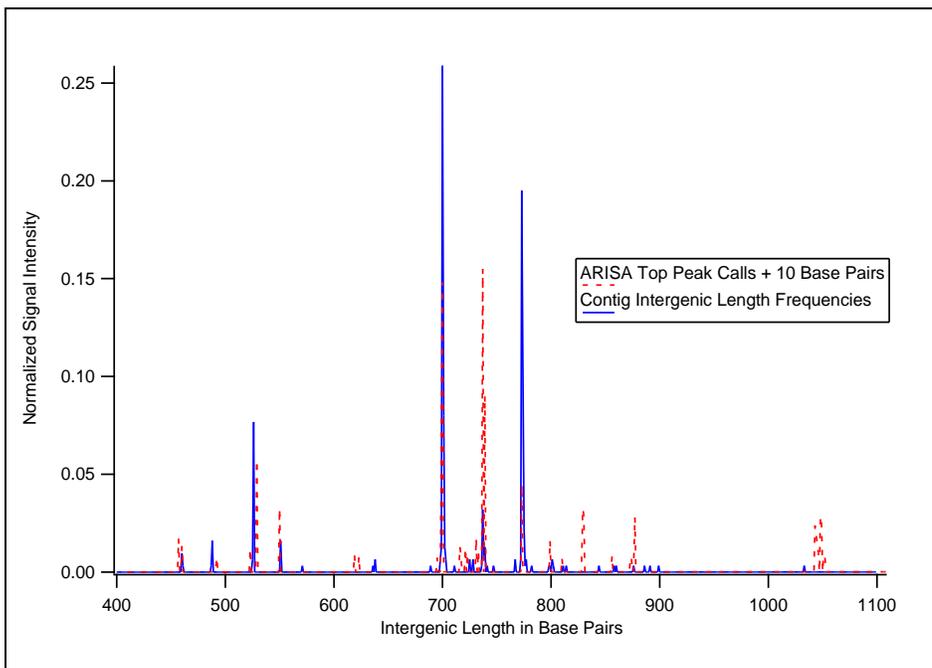


FIGURE 4.5: Comparison of normalized signal from ARISA and intergenic lengths from Sanger sequencing assemblies after 10 base pair adjustment of ARISA.

## CHAPTER 5: THE RELATIONSHIP BETWEEN ACTIVITY LEVELS AND THE GUT MICROBIAL COMMUNITY IN A MOUSE MODEL

### 5.1 Background and significance

Physical inactivity is among the leading risk-factors associated with mortality in the United States.<sup>160</sup> Mounting evidence suggests that physical activity is crucial for the health and well being of people of all ages. A sedentary lifestyle is widely known to be associated with a diverse number of health problems such as coronary heart disease and colon cancer and has been estimated to be responsible for 1 in 10 deaths in the United States each year.<sup>161</sup> Physical inactivity is also thought to promote obesity and related diseases, such as diabetes. The increasing prevalence of obesity and sedentary lifestyles in the United States<sup>162</sup> is a major public health concern, as well as a source of escalating healthcare costs. The CDC estimates that the medical costs associated with obesity are 29% to 117% greater than for people of normal weight, accounting for an estimated \$117 billion in 2000 for obesity-related health problems alone.<sup>162</sup>

Genetic differences between individuals that contribute to physical inactivity remain poorly characterized. Recent studies have observed that physical activity levels vary markedly between inbred strains of mice as measured by distance run on an activity wheel.<sup>163-164</sup> Multiple regions of chromosomes containing putative quantitative trait loci controlling various activity measures have been identified.<sup>122, 124</sup> Recently, several

specific haplotypes associated with high activity phenotypes have also been determined; however, the mechanisms by which they affect activity have not yet been discovered.<sup>165</sup>

Metagenomic research has revealed that an organism's genomic capacity extends beyond its own genome. As an example, there are ten times more microbial cells in the human gut than cells in a typical person and 100 times more microbial genes within our bodies than are in the human genome. Host and microbes together form a "super-organism" with a metagenome that acts as a single entity, sustaining the functions required for life.<sup>101</sup> As one component of this biosystem, gut microbes support a variety of metabolic processes and are critical for harvesting energy and processing nutrients from food.<sup>21, 28-29</sup> Consequently, the metabolic capacity of the superorganism is at least partially defined by these host and symbiont relationships.<sup>101</sup>

Metabolism is influenced by many factors. Results from a number of studies indicate that the gut microbiome also influences metabolic activity.<sup>21, 29, 45-46, 166-167</sup> Turnbaugh et al. have observed that gut microbiota composition varied between lean and obese individuals<sup>27</sup> and that transferring the gut microbiome from an obese mouse to a germ-free lean mouse significantly increased the weight and fat deposits of the recipient mouse.<sup>21</sup> This research led to the hypothesis that deviations from a core microbiome are responsible for the physiological alterations associated with obesity.<sup>27</sup> What is not yet clear is how the many factors that affect metabolism interact with one another to define an individual's metabolic state.

Although several studies have examined the relationship between exercise and intestinal physiology<sup>168-170</sup>, with the implicit, if not explicit, implication for involvement of the gut microbiome, research exploring the interactions between exercise and intestinal

microbes has been relatively limited.<sup>135-136</sup> Resolving questions surrounding the plasticity of an individual's gut microbiome and the degree to which controllable factors, such as physical activity levels and nutrition, can be managed are key to applying our expanding knowledge of superorganism dynamics to critical public health problems.

Testing such a complex equation requires that variables, such as diet and environment, be controlled. To investigate one of these factors, the relationship between gut microbiome composition and physical activity, we conducted three preliminary experiments using mouse models. We tested the idea that physical activity might have some effect on gut microbiome composition. Our preliminary experiments were designed to determine 1) whether we could observe any associations between exercise levels and gut microbial composition, 2) if so, the degree to which the microbial community changed in response to exercise, 3) whether there were associations between food consumption, weight gain and exercise, 4) whether small molecules in the blood were different between high activity and low activity strains and 5) which protocol choices could be applied to a full-scale experiment. Our results were planned to provide evidence to conduct further research to investigate causal relationships among these components.

We used known high activity and low activity mouse strains<sup>123-124</sup>, measured their activity levels and profiled their gut microbial communities to determine if differences in activity levels were, in fact, associated with microbiota composition differences. Each of the three preliminary experiments was designed to test a different aspect of experimental protocol and to provide evidence to develop a larger study.

All experiments for this aim were conducted in the vivarium at the University of North Carolina at Charlotte under Institutional Animal Care and Use Committee (IACUC) protocol 08-014. Microbial communities were sampled by collecting feces from each mouse and using ARISA (Chapter 4) to provide a spectral characterization of community composition from each sample. The first experiment compared activity levels and gut microbial community composition between two different mouse strains, one high activity (SM/J) and one low activity (129S1/SvImJ) strain, to determine if gut microbial community composition differed significantly between the two types of mice. In the second experiment, we used two female and two male same-strain (C57/LJ) mice to test whether changes in activity levels would result in altered gut microbiota. Our final set of experiments again used two different inbred mouse strains (C57Bl/6J and 129S1/SvImJ) to test experimental protocols for fecal sample collection, blood sample collection, food measurement and weight measurement. These measures were also analyzed with activity and microbial community assessment to look for confounding factors in our design. Blood samples from two of our mice were analyzed by mass spectrometry to identify blood metabolites that differed between the high activity and low activity mouse strains.

Peak Studio<sup>153</sup>, a set of JAVA software tools that provides comprehensive analytical capabilities for DNA molecular methods that yield spectral results, was used to analyze ARISA spectra from mouse fecal samples in Experiments 2 and 3. PeakStudio provided a platform to establish a base-pair scale for ARISA standards and data, to call data peaks from the spectra image, to assign base pair sizes to each peak and to provide summaries of binned peak call assignments for analysis.

## 5.2 Common methods

### 5.2.1 Animal care

All mice were housed individually in either 1) a standard, solid bottom, plastic rat cage with a filter top when activity wheels were not in use or 2) a large, solid bottom, plastic rat cage with a filter top equipped with an activity wheel when activity levels were being monitored. Mice were provided ad libitum standard mouse chow and water. Cages were changed every two weeks at a minimum, per IACUC recommendations. Mice were provided with fresh bedding material at each cage change. Mice were monitored either by vivarium staff or by the study team each day to ensure that they remained healthy and exhibited no signs of stress or disease. If any mouse had signs of distress, it was evaluated in consultation with the IACUC veterinary expert and a decision was made on the best way to resolve the issue.

Large rat cages were equipped with digital counters to document wheel activity. The counter consisted of a chronometer triggered by a magnet that was attached to the outside of each activity wheel. Counters were calibrated at the beginning of each experiment, at each cage change and in the event of counter or wheel malfunction. Counters were checked to document mouse wheel running activity and to ensure proper function each day.

### 5.2.2 Sample collection and preparation

For Experiments 1 and 2, and for a portion of Experiment 3, feces for metagenomic analysis were collected using sterilized forceps to remove the fecal pellet as it was expelled from the anus. We used this procedure to prevent contamination of the fecal pellet with bacteria from sources other than the gut microbiome. Each pellet was

immediately placed in a sterile labeled cryovial. When 3-5 pellets were collected, the cryovial was placed in liquid nitrogen. After the collection procedure was completed, samples were transported in liquid nitrogen to the laboratory and placed in storage at -20° C to await further preparation.

Before DNA extraction, each fecal sample was first thawed on ice. Metagenomic DNA was then extracted using the MoBio UltraClean Fecal DNA Isolation Kit (Carlsbad, CA). Extracted DNA was quantified using the Nanodrop Spectrophotometer (Wilmington, DE) to ensure sufficient DNA for analysis. DNA extracts were then stored at -20° C to await ARISA analysis.

### 5.2.3 ARISA sample preparation and analysis

The ARISA technique that was used to detect microbial signatures in our sample is based on taxonomic differences in the intergenic distance between the 16S and 23S ribosomal RNA genes (FIGURE 4.1). Using thawed, extracted DNA, the intergenic region between the 16S rRNA and the 23S rRNA genes was PCR amplified for each of two (minimum) technical replicates using a universal 16S forward primer (16S-1406F-FAM primer - 5' - /56-FAM/TGY ACA CAC CGC CCG T -3') and a eubacterial 23S reverse primer (23S-125R - 5' - GGG TTB CCC CAT TCR G -3') which was fluorescently tagged. PCR reactions were conducted using the Advantage® HD Polymerase protocol (Clontech, Mountain View, CA) with a customized master mix (TABLE 5.1). The labeled fragments were then separated by size on an Applied Biosystems 3130 sequencer (Carlsbad, CA), resulting in an electropherogram where each signal peak represented a different “operational taxonomic unit”.

## 5.3 Experiment 1: Gut metagenomics of high activity and low activity mouse strains

### 5.3.1 Experimental design

Ten mice, five each of SM/J and 129S1/SvImJ mouse strains were obtained from the lab of our collaborator, Dr. Timothy Lightfoot. The two mouse strains were selected for known and sizable differences between their activity levels.<sup>124</sup> Each mouse was placed in its own large rat cage that accommodated a standard activity wheel equipped with a digital counter to document wheel activity. The counter recorded wheel-running activity for the four-week duration of the experiment. During the study, one SM/J mouse (mouse 12) developed signs of distress and illness. In consultation with the vivarium veterinarian, a decision was made to euthanize the animal, and its samples and activity observations were excluded from the analysis. Methods of preparing fecal samples for ARISA are described in Common Methods.

### 5.3.2 ARISA analysis

The electropherograms from ARISA were analyzed using a preliminary version of custom JAVA software that applied size standards to ARISA spectra, identified peaks in the spectra and assigned base pair lengths to each called peak. The vectors resulting from peak assignment were compared to one another to assess similarities and differences. The R statistical software package was used to conduct hierarchical clustering of sample vectors that represented each ARISA spectrum.

### 5.3.3 Results

We observed large and sustained disparities in activity levels (measured by daily distance run on an activity wheel) between SM/J and 129 mouse strains (FIGURE 5.1). A comparison of ARISA signals from the fecal samples showed that there were also

differences in gut microbiota profiles between the two mouse strains. We used hierarchical clustering to analyze the ARISA spectra and visualize the comparison. All samples clustered perfectly by mouse strain, indicating clear differences between the strains in overall gut microbial composition (FIGURE 5.2). Although our results demonstrated that the high active and low active mouse strains each had distinctive gut microbial profiles, the design of the experiment did not allow us to separate the effects from strain genomics, from disparities in gut microbiome acquisition (either from dam or sibling exposure) or from activity level differences.

#### 5.4 Experiment 2: Effect of activity changes on mouse gut metagenomics

##### 5.4.1 Experimental design

Experiment 2 was designed to test the effect of activity levels on mouse gut microbial composition. Unlike in Experiment 1, in which we compared two different strains with dissimilar activity levels, this experiment used four C57/LJ high activity mice, 2 female and 2 male, provided by the lab of Dr. Timothy Lightfoot. To control for potential effects from differences in gut microbiome acquisition, we used mouse siblings and employed a cross-over design. This design allowed each mouse to act as its own control, supplying a built-in comparison of each set of experimental conditions. In Phase 1, the mice were provided *ad libitum* access to an activity wheel for 2 weeks, they then had the wheel removed for a period of time for 2 weeks in Phase 2 and finally had the wheel returned to their cages for another 2 weeks for Phase 3 (FIGURE 5.3). Sampling was conducted on a regular schedule through all three phases of the experiment resulting in a total of 147 samples, including technical replicates. Methods of preparing fecal samples for ARISA are described in Common Methods.

#### 5.4.2 ARISA analysis

A minimum of two technical replicates of extracted 16S rRNA gene intergenic sequence from each sample were run using ARISA. The resulting electropherograms were analyzed using Peak Studio.<sup>153</sup> For each of the three experimental phases, the last four fecal samples obtained during that phase from each mouse were selected for analysis. These samples were determined to best represent the gut microbiota based on the conditions during that phase, as the animals had the opportunity to adjust to changed conditions. A minimum of two technical replicates of ARISA spectra were examined. One of these was selected for further analysis based on the consistency and quality of the standard peak signal, which is used to set the base pair scale, and of the data peak signal.

Spectral peaks were identified and called based on binned base pairs (bin size = 3)<sup>171</sup> for the range from 250 to 1000 base pairs. Each analyzed sample resulted in an array in which each data value for a bin represented the signal intensity of the peak call for that bin, if any peak were called. The arrays were consolidated into a matrix by bin size for further analysis.

#### 5.4.3 Statistical analysis

The matrix of binned peak call arrays that included the last four samples from each phase (n=48) was then used to assess the similarities and differences between samples with regard to gut microbiota profiles. Hierarchical clustering used Ward's method on binned signal intensities standardized to the average signal intensity for all arrays in the sample, as did Principal Components Analysis on covariance. The following formula was used to standardize peak intensities for each sample,

$$\left( \frac{S_i}{\sum_{i=1}^n S_i} \right) * \mu_{S_i}$$

where  $S_i$  is the signal intensity value in each bin.

Statistical analyses were conducted and figures constructed using JMP® 8.0 software for Microsoft Windows (SAS Institute) and Microsoft® Office Excel 2003.

#### 5.4.4 Results

In Phases 1 and 3 of our experiment, all mice were exposed to an activity wheel and running distance was measured each day. As expected, the C57/LJ mouse strain exhibited high levels of activity (FIGURE 5.4), evidenced by daily running distance (mean = 8.76 km/day; Phase 1 mean = 10.42 km/day; Phase 2 mean = 7.10 km/day). Principal Components Analysis of running distances showed differences between Phase 1 and Phase 3 (FIGURE 5.5). All mice demonstrated significantly lower Phase 3 activity levels (FIGURE 5.6). Mice also exhibited differences from each other in their activity levels (FIGURE 5.7). When Phase 1 and Phase 3 were analyzed separately, differences between mice were more apparent in Phase 3.

Mouse weights were measured throughout the study. All four mice gained weight from the beginning to the end of the experiment (FIGURE 5.8). Weight differences between mice appeared to be associated with sex; the male mice consistently weighed more than the female mice. Interestingly, weights for all mice dropped immediately after activity wheels were removed from the cages for Phase 2 (mean = 4.22%). No mice regained the weight quickly and all but one mouse, 340-1, did not regain the lost weight until the activity wheels were returned to the cages in Phase 3.

ARISA was used to profile the gut microbiota from mouse fecal samples over the course of the experiment. ARISA analysis was completed on extracted DNA from same-day baseline fecal samples from each of the four mice (FIGURE 5.9). The superimposed spectra show that the mice had very similar microbiome profiles at baseline.

To compare the three different experimental phases, the last four samples taken during each phase were selected for analysis to ensure that microbial composition had stabilized and that it would reflect the metabolic characteristics of that phase. Hierarchical clustering of the arrays of standardized peak intensities resulting from the analysis of ARISA spectra demonstrated differences between gut microbiota in each of the phases of the experiment (FIGURE 5.10). The most consistent clustering among the samples occurred for Phase 2, the period where activity was restricted. Only 2 samples from Phase 2 did not cluster perfectly, both from mouse 338-1, the mouse with the lowest mean running distances in both Phases 1 and 3. When we conducted Principal Components Analysis on covariances of the ARISA results, the size of differences between phases was apparent, with Phase 2, the period during which activity levels were restricted, showing the greatest separation from the phases during which activity wheels were available (FIGURE 5.11). After assessing sample distribution normality and variance equivalence, we tested the first component of the PCA to determine whether there were differences in means between samples taken when mice were running and those that were taken when activity wheels were not available, Phases 1 & 3 vs. Phase 2 (FIGURE 5.12). The Wilcoxon test indicated that the gut microbiota between phases were different ( $P = 1.5e-6$ ), indicating that lack of activity was a significant factor in differentiating the mouse gut microbiota. It is also noteworthy that the Jackknife

Distance method identified a few observations in each distribution that could be classified as outliers, the majority of which were samples from mouse 338-1.

#### 5.4.5 Discussion

Health problems from increasing levels of obesity and sedentary lifestyles have been widely documented.<sup>121, 137-138</sup> Several studies have investigated differences in activity levels between mouse strains in an attempt to identify factors, such as host genetics, that might account for activity disparities.<sup>164-165, 172-175</sup> The latest in this research has documented gene, as well as non-coding region, associations that correspond to activity differences and suggest moderate to high heritability.<sup>165</sup> These studies offer important contributions to improve understanding of innate differences in exercise propensity but provide less information on the mechanisms that govern the relationship between exercise and metabolism, an essential component for improving health outcomes.

Although ground-breaking studies have shown an association between obesity, metabolism and the composition of the gut microbiome<sup>21, 27, 29, 46-48, 167, 176</sup> very little research has investigated the relationship between gut microbial composition and exercise. One human study examined the effects on the gut microbiota of obese adolescents of weight loss from diet and exercise.<sup>136</sup> Using qPCR, their results revealed changes to specific bacterial groups but did not determine which factors led to the microbial composition change. Matsumoto et al. provided more direct evidence that exercise alters the gut microbiome.<sup>177</sup> Their study of voluntary exercise in rats used cecal samples to compare qPCR-TGGE profiles of rats exposed to activity wheels to a control group which were not provided wheel running opportunities. Similar to our results, hierarchical clustering revealed that the majority of samples clustered by exercise group.

Because the Matsumoto study used a cohort design and collected only samples that required animal sacrifice, the analysis could not identify exercise-related gut microbiome changes within each animal, unlike the research reported here.

Our results revealed gut microbial composition change in each mouse that was significantly associated with activity levels (FIGURE 5.12). For the 8 observations that did not follow the common pattern (the outliers) in the means test of PC1 against activity phases, all but one belonged to mouse 338-1, the mouse with the lowest average running distance, reinforcing the interpretation that activity levels affect gut microbes. The obvious explanation of these results, that exercise affects metabolism, is unsurprising. The further implications that gut microbes are intricately bound to host metabolic changes and that microbial composition is alterable are more significant and underscore the importance of further investigation to unravel the mechanisms of host and microbiome symbiosis that are involved in metabolism.

The results from Experiment 2 allowed us to examine the effects of changing activity wheel availability, a proxy for mouse activity levels, on gut microbial community composition profiles. Using a cross-over design provided a method to manage confounding factors for a preliminary study using a small number of animals. While these results provide important clues to generate hypotheses for further studies, ultimately, the advantages of the experimental design alone were not enough to overcome the limitations of small sample size, and our results, while interesting, must still be considered preliminary.

## 5.5 Experiment 3: Protocol testing for mouse activity studies

### 5.5.1 Experimental design

Experiment 3 was designed to test procedures and protocols that were planned for use in a larger study of gut microbiome and activity levels. Twelve female mice, six each of 129S1/SvImJ low activity and C57BL/6J high activity strains, were purchased from The Jackson Laboratory (Bar Harbor, ME) at six weeks of age. Mice were quarantined for 3 days per laboratory protocols before being placed in individual cages with running wheels.

We recognized that, for our study to have sufficient power, a much larger sample size would be needed and that some of our existing methods were not practically scalable. As an example, our existing protocol that used forceps to collect fecal samples required between 5-20 minutes per mouse per sample. For this experiment, we tested the feasibility of a more efficient sample collection technique where we simultaneously placing the mice in sterilized, individual, empty rat cages, allowing them time to defecate and collecting the samples from the cage. Our design for a larger study also included an assessment of associations between small molecules in the blood and gut microbiome composition. Therefore, we tested methods of collecting blood samples. Finally, we were interested in assessing the potential differences in microbial community colonization between cecal, intestinal and fecal microbiota<sup>46, 178</sup> to inform our decisions for future gut microbiome sampling protocols.

### 5.5.2 Fecal sample collection

Clean, empty large rat cages were washed with disinfecting solution and dried. Each mouse was placed in a cage by itself. Once 3-5 fecal pellets were produced, the mouse

was removed and returned to its previous cage. Feces were then collected from the empty cage with sterile forceps and put in a labeled cryovial which was placed in liquid nitrogen. Samples were then frozen at -20° C. Methods of preparing fecal, cecal and intestinal samples for ARISA are described in Common Methods.

### 5.5.3 ARISA analysis

A minimum of two technical replicates of extracted 16S rRNA intergenic sequence from each sample were run using ARISA. The resulting electropherograms were analyzed using Peak Studio.<sup>153</sup> Spectral peaks were identified and called based on binned base pairs (bin size = 3)<sup>171</sup> for the range from 250 to 1000 base pairs. Each analyzed sample resulted in an array in which each data value for a bin represented the signal intensity of the peak call for that bin, if any peak were called. For comparison of cecal, fecal and intestinal samples, those sample arrays were consolidated into a matrix by bin size for further analysis.

### 5.5.4 Statistical analysis

The matrix of binned peak call arrays was then used to assess the similarities and differences between samples with regard to gut microbiota profiles. Hierarchical clustering used Ward's method on binned signal intensities standardized to the average signal intensity for all arrays in the sample, as did Principal Components Analysis on covariance. Peak intensities were standardized for each sample as previously described.

Statistical analyses were conducted and figures constructed using JMP® 8.0 software for Microsoft Windows (SAS Institute) and Microsoft® Office Excel 2003.

#### 5.5.5 Blood sample collection

Study personnel were trained in survival surgery, blood sampling and anesthesia using isoflurane per IUCAC protocol. Blood sample volume maximums were equal to the lesser of 150-200ul or < 10% of blood volume calculated from mouse body weight and were collected from the saphenous vein with mice under isoflurane anesthesia using a heparinized collection tubes. Actual blood obtained from two survival surgery collections was of insufficient volume (20-100ul whole blood per mouse) to yield enough serum for mass spectrometry. Therefore, two mice (129-5 and C57-4) underwent anesthesia and non-survival surgery for blood collection through posterior vena cava puncture. Collected blood was then transferred to labeled vials and immediately transported to the laboratory for serum preparation.

#### 5.5.6 Metabolite extraction

50- $\mu$ L aliquot of serum sample was spiked with an internal standard solutions (10 $\mu$ L heptadecanoic acid in methanol, 1 mg/mL) and vortexed for 10 seconds. The mixed solution was extracted with 150  $\mu$ L of pre-cooled (-20°C) methanol: chloroform (3:1) and vortexed for 30 seconds. After storing for 10 minutes at -20°C, the samples were centrifuged at 10,000 g for 10 minutes at 4°C. An aliquot of the 170- $\mu$ L supernatant was transferred to a glass sampling vial to vacuum dry at room temperature.

#### 5.5.7 Chemical derivatization

The residue was derivatized using a two-step procedure. First, 80  $\mu$ L methoxyamine (15 mg/mL in pyridine) was added to the vial and kept at 30°C for 90 minutes. An amount of 10  $\mu$ L retention index compounds (the mixture of C10-C40, 50 $\mu$ g/mL), and 80

$\mu\text{L}$  BSTFA (1%TMCS) were added into the reaction vials. Then the samples were subjected to a 70°C for 120 minutes derivatization reaction.

#### 5.5.8 Gas chromatography time-of-flight mass spectrum (GC-TOFMS) analysis

Each 1- $\mu\text{L}$  aliquot of the derivatized solution was injected in splitless mode into an Agilent 7890N gas chromatograph coupled with a Pegasus HT time-of-flight mass spectrometer (Leco Corporation, St Joseph, USA). Separation was achieved on a DB-5 ms capillary column (30 m  $\times$  250  $\mu\text{m}$  I.D., 0.25- $\mu\text{m}$  film thickness; Agilent J&W Scientific, Folsom, CA, USA), with helium as the carrier gas at a constant flow rate of 1.0 ml/min. The temperature of injection, transfer interface, and ion source was set to 260°C, 260°C, and 210°C, respectively. The GC temperature programming was set to 2 min isothermal heating at 80°C, followed by 10°C/min oven temperature ramps to 220 °C, 5 °C/min to 240°C, and 25°C/min to 290 °C, and a final 8 min maintenance at 290°C. Electron impact ionization (70 eV) at full scan mode (m/z 40-600) was used, with an acquisition rate of 20 spectra/second in the TOFMS setting.

#### 5.5.9 GC-TOFMS data processing

The data generated in GC-TOFMS instrument were analyzed by ChemTOF software (v4.22, Leco Co., CA, USA). Peak areas of unique mass were normalized to the internal standard. Metabolites were identified by comparing NIST library and confirmed by our established library.

#### 5.5.10 Results

In Experiment 3, we tested sampling methods and protocols, as well as collecting activity levels, weights and food consumption. In two of our mice, C57BL/6J (high activity) mouse and a 129S1/SvImJ (low activity) mouse, we also analyzed ARISA gut

microbiota profiles from cecal, fecal and intestinal samples and metabolomic profiles of serum samples. For these two mice, documentation of activity levels, measured as distance run on an activity wheel confirmed that running distance for the C57 individual was consistently 10-fold or higher than for the 129 individual over a period of several weeks. Daily food consumption of the C57 mouse was also consistently higher than that of the 129 mouse.

A comparison cecal, fecal and intestinal sample profiles using hierarchical clustering on ARISA binned signals for mice 129-5 and C57-4 (FIGURE 5.13) showed that each mouse's samples clustered together, indicating that differences between mice were larger than differences between sample sources. The findings from Principal Components Analysis on covariances were less conclusive, however. Although the three samples from mouse 129-5 clustered tightly, those from C57-4 demonstrated much more separation between the three sources.

Preliminary results of metabolomic analysis from serum samples for 129-5 and C57-4 yielded ~450 peaks of which ~180 match reference metabolites from our established library. Peak areas of unique mass were normalized to the internal standard. Metabolites were identified via comparison to our established library.

The total ion chromatograms (TIC) of the two samples are shown in FIGURE 5.14. As might be expected given that the mice were in the common environment of the UNC Charlotte Animal Facility and had been exposed to the same diet of mouse chow, there were considerable overlaps between the two spectra. There were also, however, several clear differences in the relative magnitudes of some of the peaks. In order to quantify these differences, we used the ChemTOF software to call 453 distinct peaks in the sample

from the 129-5 mouse and 449 peaks in the sample from the C57-4 mouse. Of these ~450 peaks, around 180 matched metabolites that we had identified in our reference library.

In order to examine metabolites that might be differentially expressed between our two samples, we chose an arbitrary cutoff of a two-fold change. By this threshold, we observed 53 differential peaks, of which we could identify 22 (TABLE 5.2). Obviously, with only  $n=1$  in each category, we do not know the variance of each peak and therefore can not perform meaningful statistics to evaluate for each of the metabolites the statistical significance of a null hypothesis of no differential expression. Nonetheless, these preliminary data are encouraging as they demonstrated 1) that we have the technical capability to achieve the metabolomics experiments described in the proposal and 2) that there are likely a non-trivial number of differentially expressed metabolites between high-active and low-active strains.

TABLE 5.1: MasterMix formulation for PCR reactions of intergenic spacer region between 16S and 23S rRNA genes.

Reagents	Volume/Final Concentrations
Advantage HD buffer (Mg <sup>2+</sup> )	5 µl
dNTP Mixture (10mM each)	0.5 µl
Advantage HD Polymerase (2.5 units/µl)	0.25 µl
16S-1406F-FAM primer 5' - /56-FAM/TGY ACA CAC CGC CCG T -3'	2.5 µl
23S-125R 5' - GGG TTB CCC CAT TCR G -3'	2.5 µl
H <sub>2</sub> O	9.25 µl
Total	20 µl MM each rxn. + 5 µl template each rxn.

TABLE 5.2: Differentially expressed metabolites from serum samples analyzed using mass spectrometry.

Metabolite	Fold change (<1 indicates higher in 129)
Propanoic acid, 2,3-bishydroxy	-50.37
d-Galactose	-5.76
Glycine	-4.53
Cholesterol	-3.96
Á-D-Galactofuranose,	-3.81
L-Ascorbic acid	-3.78
1-Pentamethyldisilyloxyptane	-3.60
Butanoic acid, 3-methyl-3-hydroxy-	-3.21
Urea	-3.13
Oleic acid	-2.79
Butanedioic acid	-2.74
Myo-Inositol, phosphate	-2.72
Glucopyranose	-2.53
3-hydroxycaproate	-2.34
Arabitol	2.10
Pyruvic acid	2.19
2-Butenedioic acid (E)-	2.25
Inositol	2.56
d-Glucose	3.82
Hydroxy acetic acid	4.94
Phosphate	5.83
Citrulline	10.39

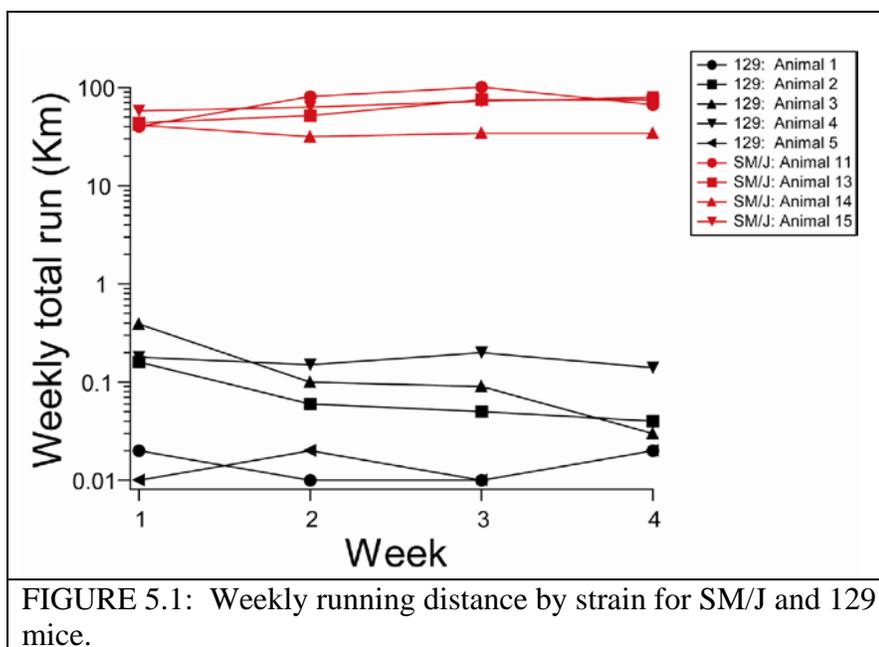


FIGURE 5.1: Weekly running distance by strain for SM/J and 129 mice.

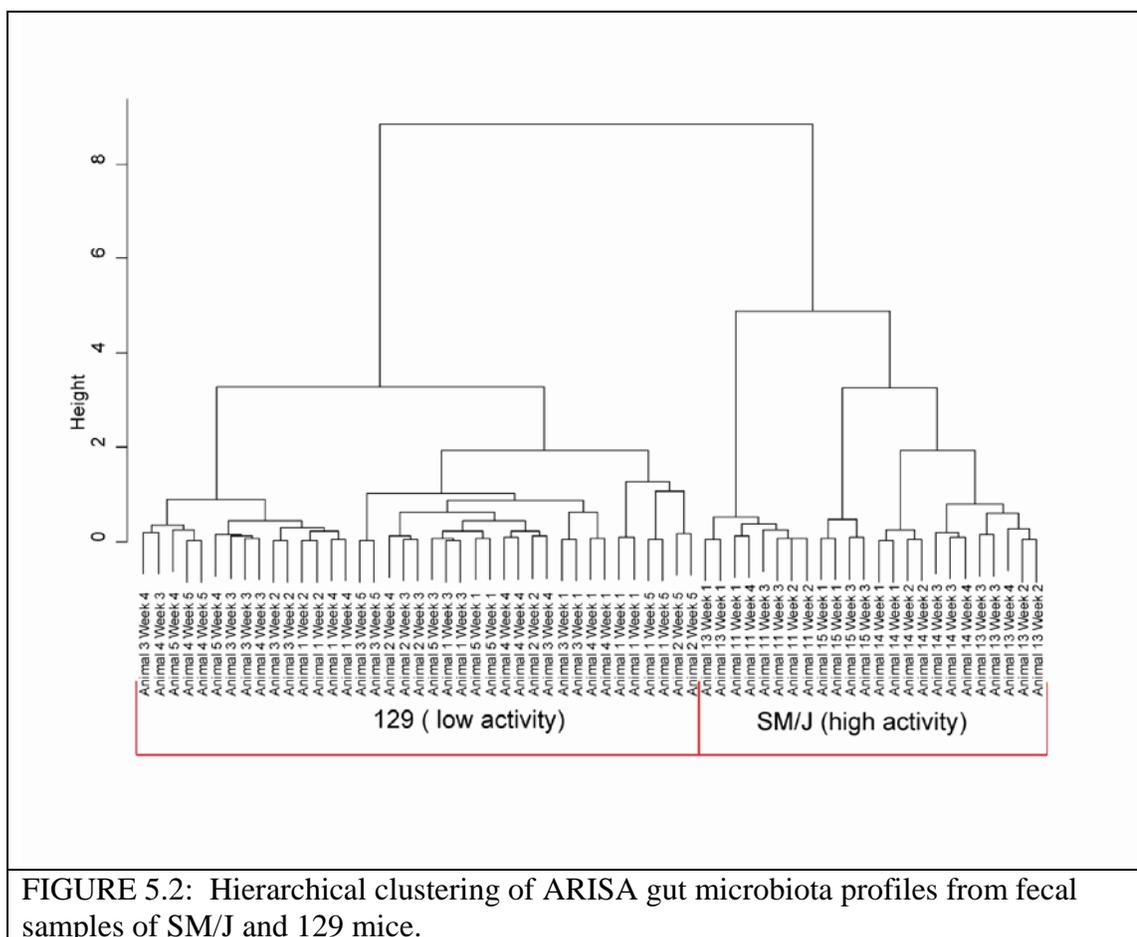
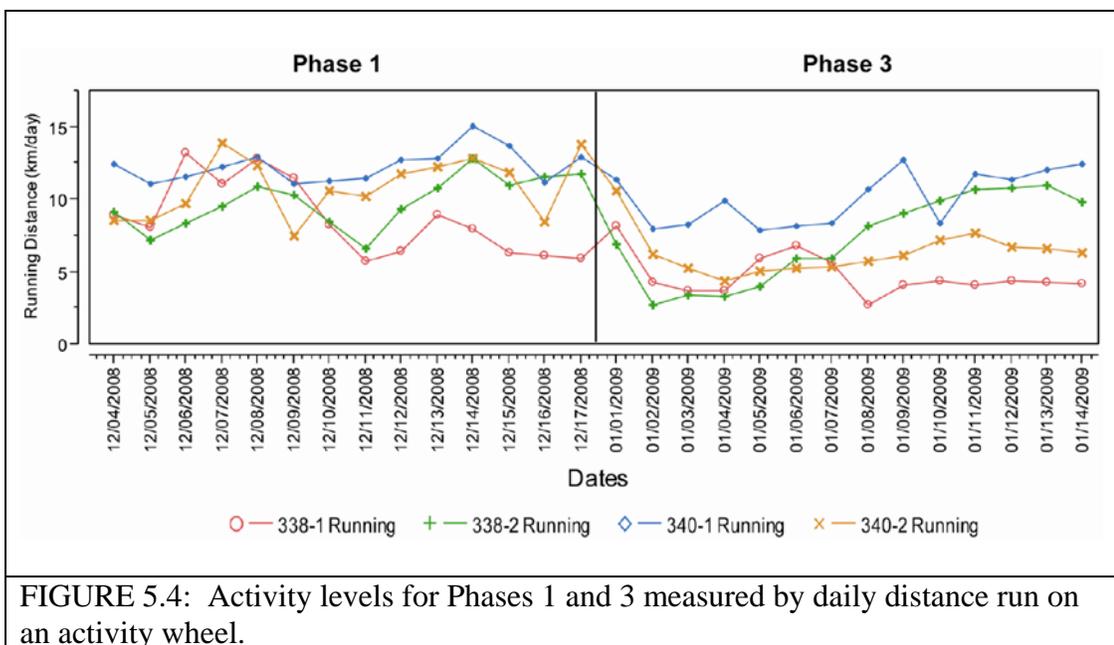
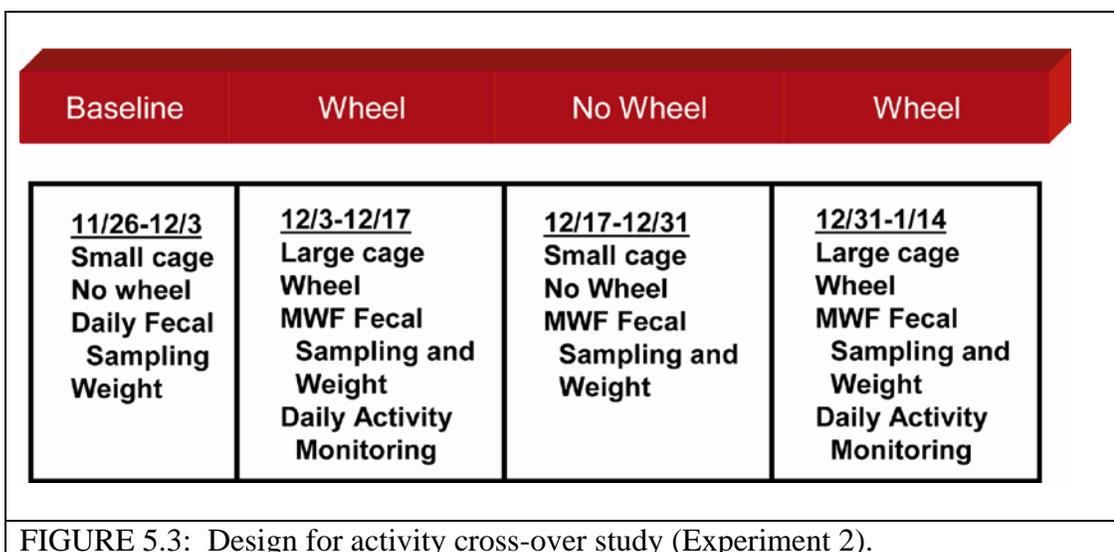


FIGURE 5.2: Hierarchical clustering of ARISA gut microbiota profiles from fecal samples of SM/J and 129 mice.



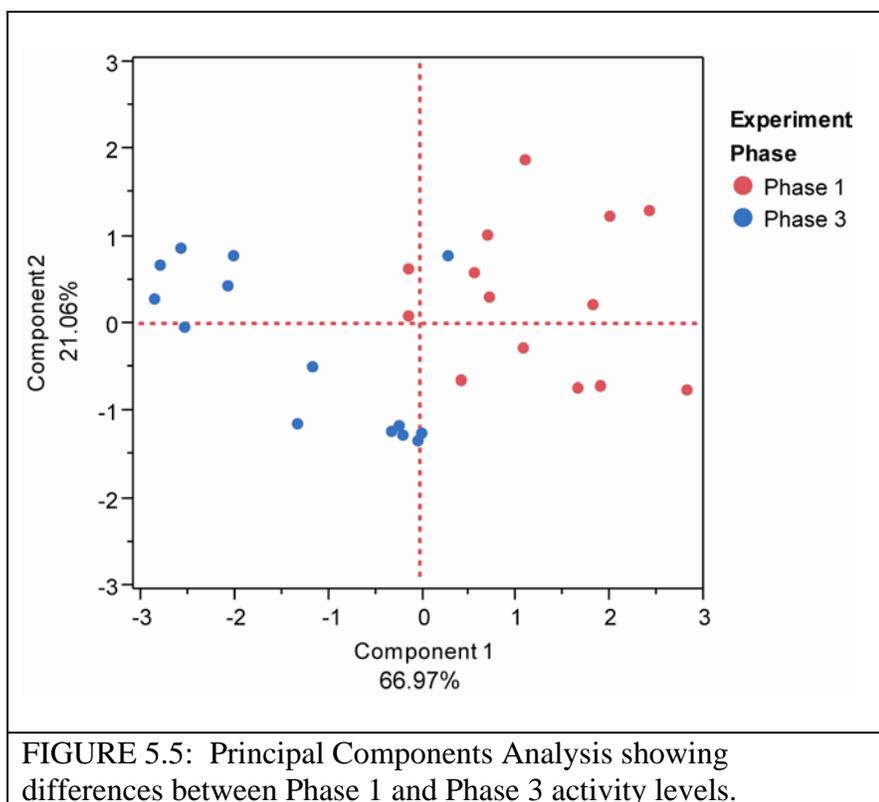


FIGURE 5.5: Principal Components Analysis showing differences between Phase 1 and Phase 3 activity levels.

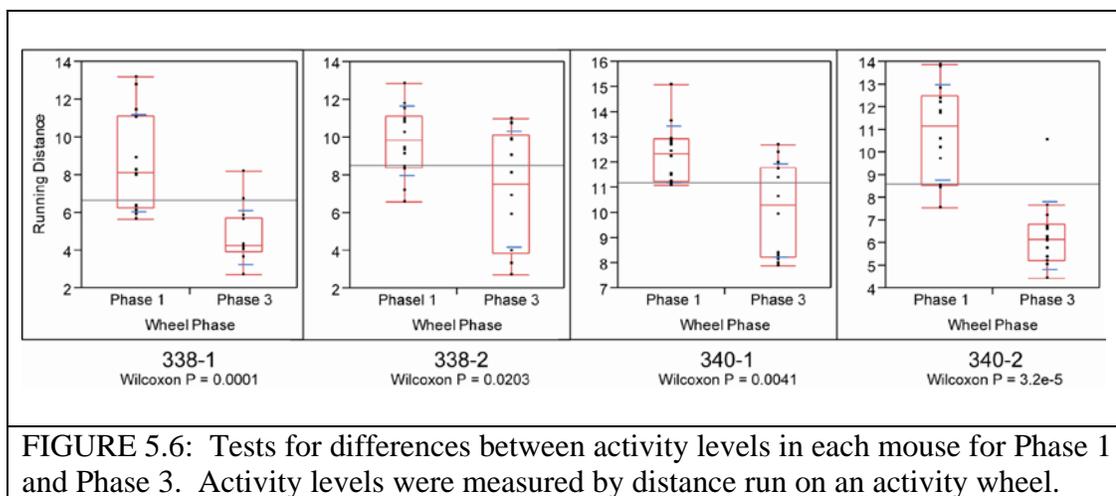
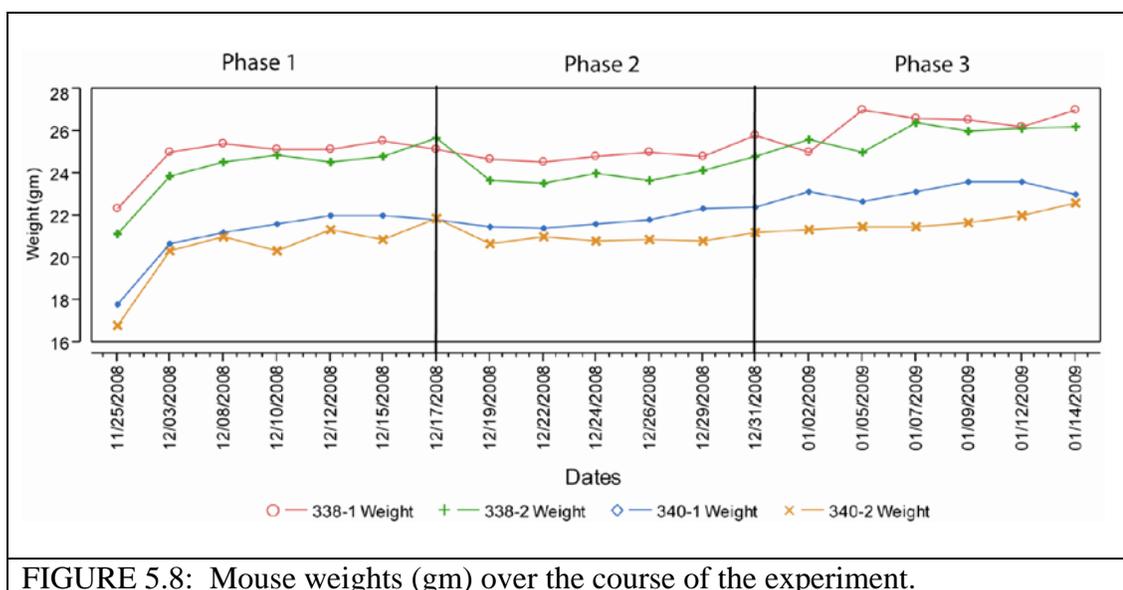
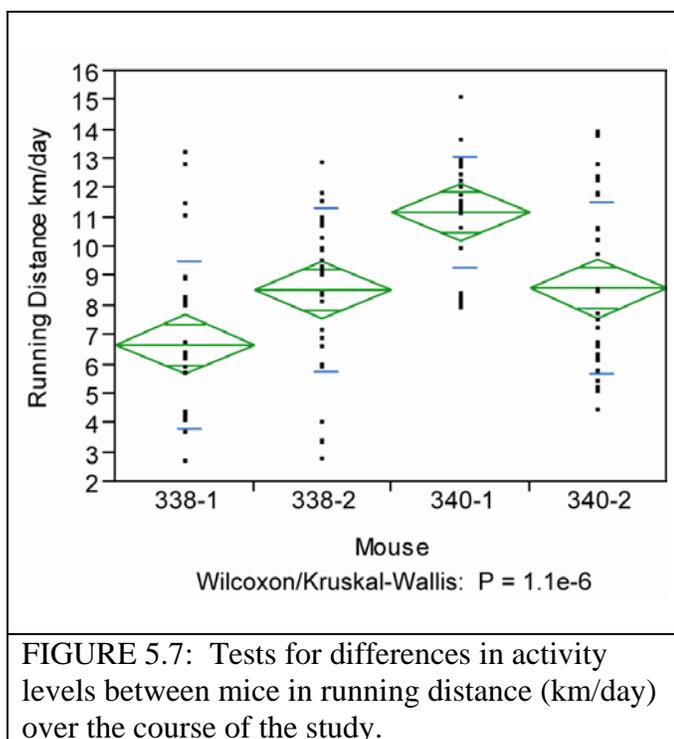
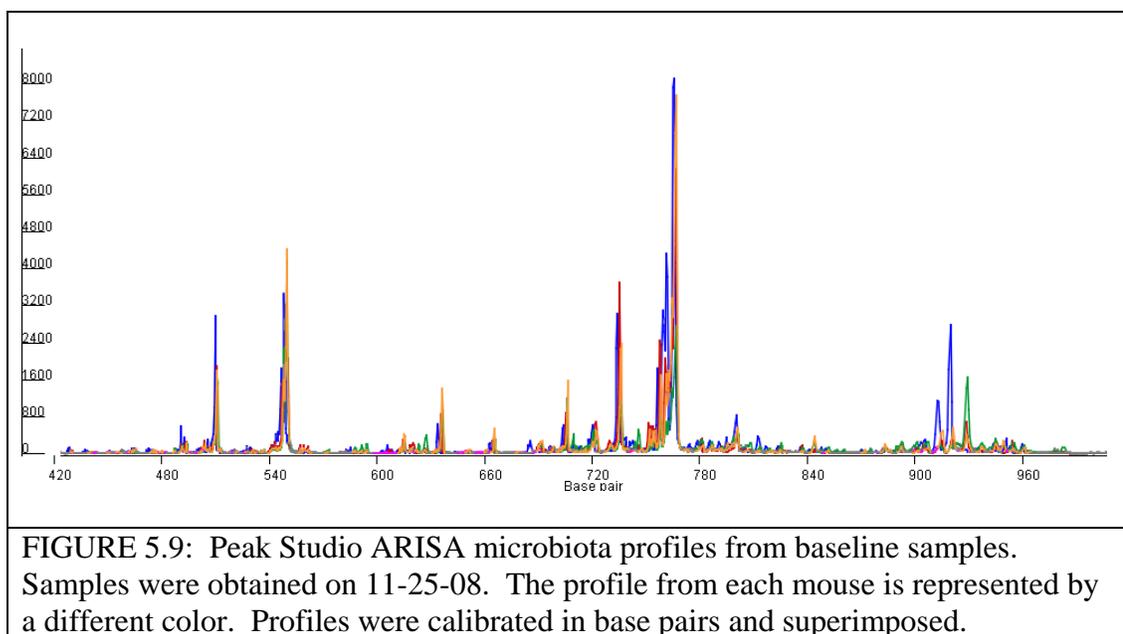


FIGURE 5.6: Tests for differences between activity levels in each mouse for Phase 1 and Phase 3. Activity levels were measured by distance run on an activity wheel.





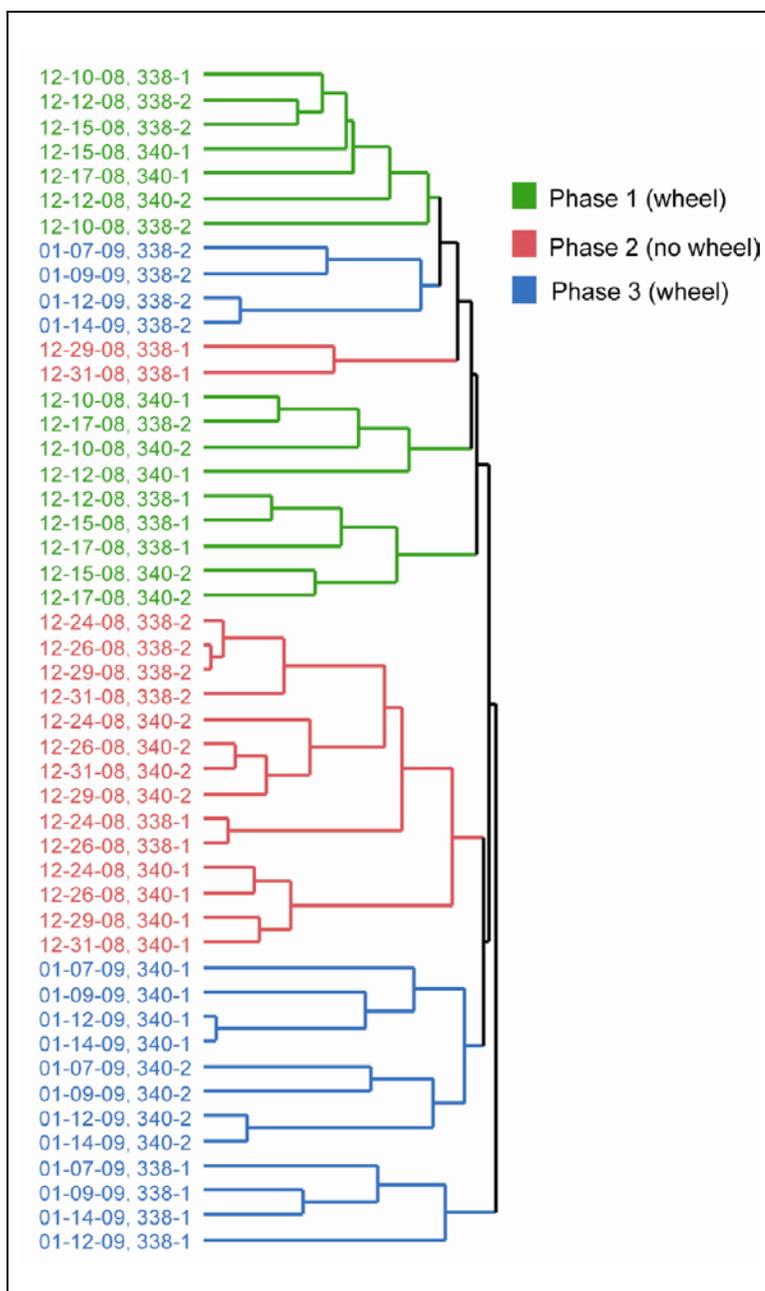


FIGURE 5.10: Hierarchical clustering of ARISA gut microbiome profiles from mouse fecal samples. The last four samples from each phase were analyzed. Sample labeling indicates sampling date and mouse number. Green = Phase 1 where activity wheel was provided. Red = Phase 2 where mice were in small cages with no activity wheel. Blue = Phase 3 where activity wheels were re-introduced.

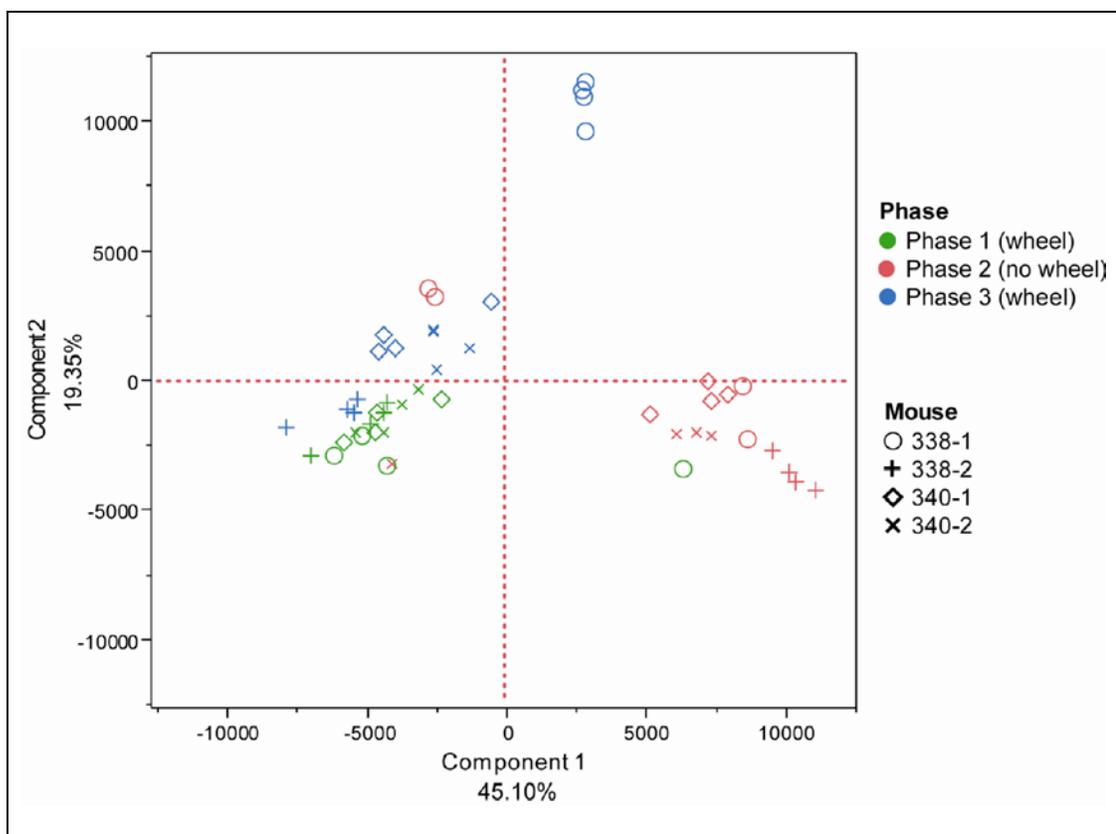


FIGURE 5.11: Principal Components Analysis on covariances of gut microbiome profiles from binned ARISA signals. The plot shows separation of the profiles from phases where activity wheels were available (Phase 1 = green, Phase 3 = blue) and the phase during which activity was restricted (Phase 2). Profiles from each mouse are designated by different symbols.

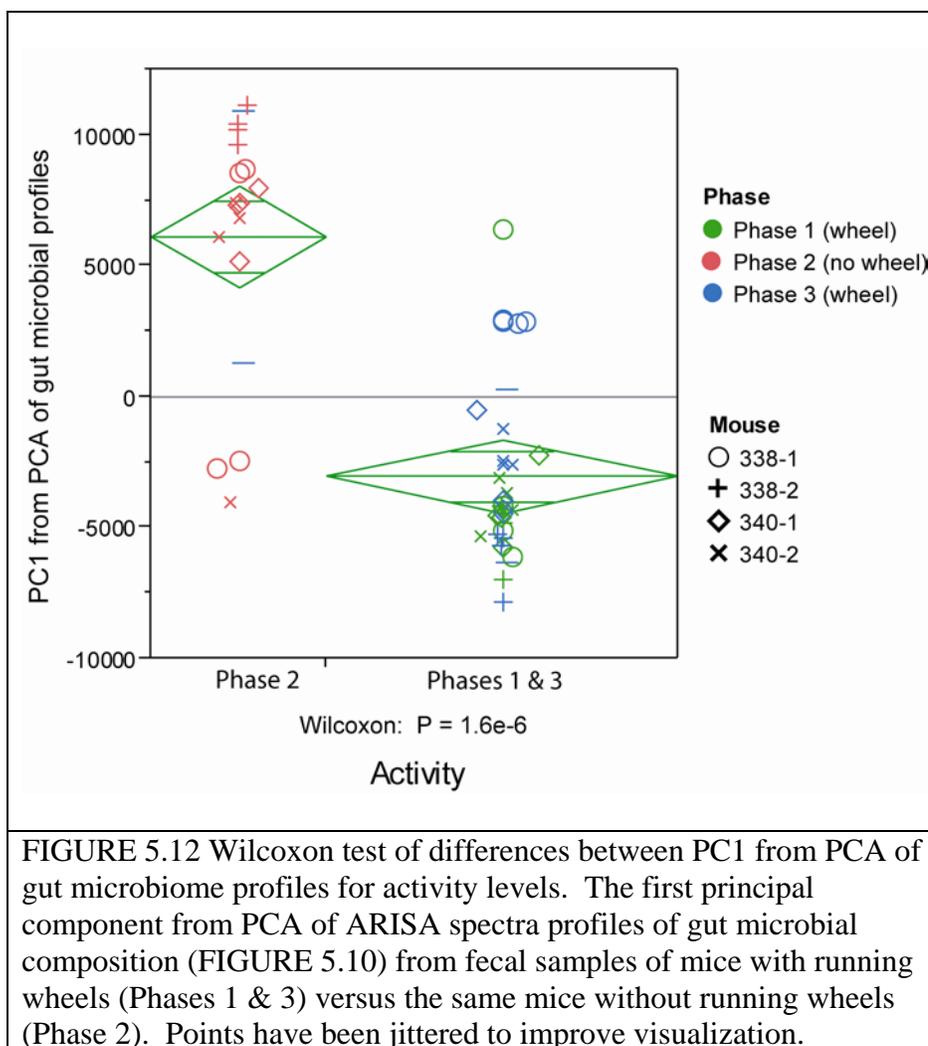


FIGURE 5.12 Wilcoxon test of differences between PC1 from PCA of gut microbiome profiles for activity levels. The first principal component from PCA of ARISA spectra profiles of gut microbial composition (FIGURE 5.10) from fecal samples of mice with running wheels (Phases 1 & 3) versus the same mice without running wheels (Phase 2). Points have been jittered to improve visualization.

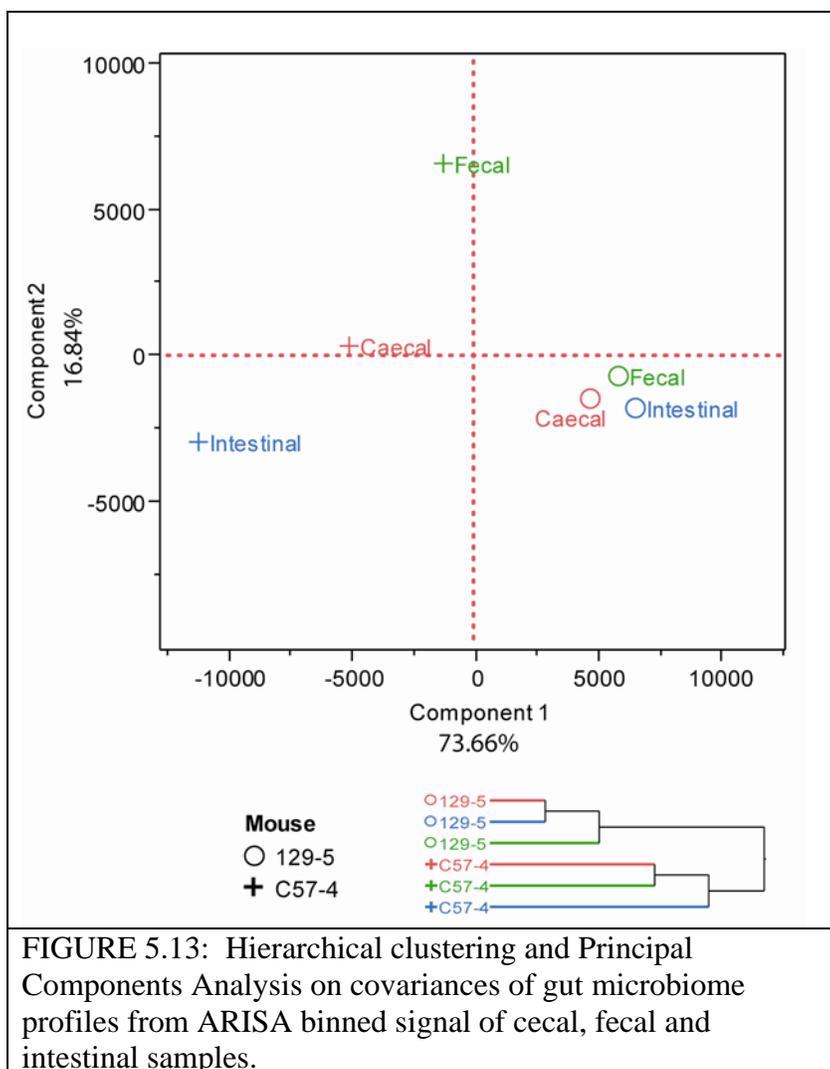
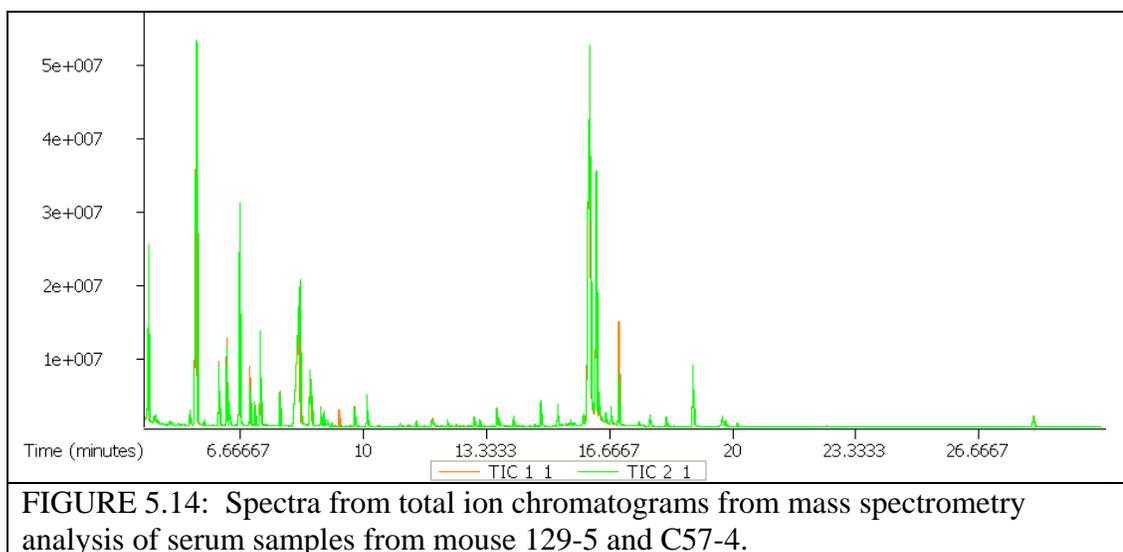


FIGURE 5.13: Hierarchical clustering and Principal Components Analysis on covariances of gut microbiome profiles from ARISA binned signal of caecal, fecal and intestinal samples.



## CHAPTER 6: CONCLUSIONS

The metagenomic studies described in this dissertation illustrate dynamic relationships among microbiota, host factors and environmental factors that define mammalian biosystems. These results relied on advances in DNA sequencing technology to characterize microbiome composition and its associations with host health. Low resolution molecular techniques, such as ARISA, that were state-of-the-art at the beginning of this research, were quickly supplanted by the highly detailed microbiome profiles provided by next-generation sequencing methods. This technology revolution has created tremendous opportunities to expand our understanding of host-microbiome interactions but has also driven a proliferation of bioinformatics and analytical methods that have complicated the quest for biologically relevant results.

In this environment of rapid change, technical and analytical choices have increased in importance. The initial selection of a sequencing platform requires understanding the trade-offs between the volume of sequences produced, sequence length, error rates, ease of sample identification, reliability, speed and expense. The choice of a data organization scheme demands that algorithm details are well-understood and that the implications of different grouping or classification methods are appreciated. Correctly identifying statistically significant results necessitates that hypothesis testing is planned in advance, that the most appropriate tests are used and that the possibility of spurious associations is

minimized through correction. Nevertheless, the most carefully selected techniques are no protection against unanticipated methodological artifacts, like those uncovered and addressed in Chapter 3.

Such circumstances make it easy to get lost in the intricacies of technology and methodology decisions, especially when every new advance has its own set of advocates and few precedents have been established. Furthermore, the excitement of embracing the next new technology can obscure the relationship between biological question and method choice. In some cases, that relationship is critically important and can even reveal that an old-school approach can augment a deficiency in a newer technique. As this dissertation clearly demonstrates, focusing on the biological questions is imperative and provides both barometer for determining biological relevance and a roadmap for navigating the complex decisions inherent in metagenomic research.

## REFERENCES

1. Amann R, Ludwig W, Schleifer K. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev* 1995;59:143-169.
2. Rusch DB, Halpern AL, Sutton G, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 2007;5:e77.
3. Sogin ML, Morrison HG, Huber JA, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 2006;103:12115-20.
4. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. *Nature* 2007;449:804-10.
5. Palmer C, Bik EM, DiGiulio DB, et al. Development of the human infant intestinal microbiota. *PLoS Biol* 2007;5:e177.
6. Dantas G, Sommer MO, Oluwasegun RD, et al. Bacteria subsisting on antibiotics. *Science* 2008;320:100-3.
7. Lear G, Niyogi D, Harding J, et al. Biofilm bacterial community structure in streams affected by acid mine drainage. *Appl Environ Microbiol* 2009;75:3455-60.
8. Sanapareddy N, Hamp TJ, Gonzalez LC, et al. Molecular diversity of a North Carolina wastewater treatment plant as revealed by pyrosequencing. *Appl Environ Microbiol* 2009;75:1688-96.
9. Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time. *Science* 2009;326:1694-7.
10. Dumas ME, Barton RH, Toye A, et al. Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. *Proc Natl Acad Sci U S A* 2006;103:12511-6.
11. Toye AA, Dumas ME, Blancher C, et al. Subtle metabolic and liver gene transcriptional changes underlie diet-induced fatty liver susceptibility in insulin-resistant mice. *Diabetologia* 2007;50:1867-79.
12. Servin AL. Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiol Rev* 2004;28:405-40.
13. Brook I. The role of bacterial interference in otitis, sinusitis and tonsillitis. *Otolaryngol Head Neck Surg* 2005;133:139-46.

14. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;437:376-80.
15. Spencer MD, Hamp TJ, Reid RW, et al. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology* 2010.
16. Wahab AA, Janahi IA, Marafia MM, et al. Microbiological identification in cystic fibrosis patients with CFTR I1234V mutation. *J Trop Pediatr* 2004;50:229-33.
17. Harrison F. Microbial ecology of the cystic fibrosis lung. *Microbiology* 2007;153:917-23.
18. Sibley CD, Rabin H, Surette MG. Cystic fibrosis: a polymicrobial infectious disease. *Future Microbiol* 2006;1:53-61.
19. Doud MS, Light M, Gonzalez G, et al. Combination of 16S rRNA variable regions provides a detailed analysis of bacterial community dynamics in the lungs of cystic fibrosis patients. *Hum Genomics* 2010;4:147-69.
20. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;11:733-9.
21. Turnbaugh PJ, Ley RE, Mahowald MA, et al. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 2006;444:1027-31.
22. Backhed F, Manchester JK, Semenkovich CF, et al. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proc Natl Acad Sci U S A* 2007;104:979-84.
23. Vijay-Kumar M, Aitken JD, Carvalho FA, et al. Metabolic Syndrome and Altered Gut Microbiota in Mice Lacking Toll-Like Receptor 5. *Science* 2010.
24. Jansson J, Willing B, Lucio M, et al. Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One* 2009;4:e6386.
25. Fox JG, Feng Y, Theve EJ, et al. Gut microbes define liver cancer risk in mice exposed to chemical and viral transgenic hepatocarcinogens. *Gut* 2010;59:88-97.
26. Davis CD, Milner JA. Gastrointestinal microflora, food components and colon cancer prevention. *J Nutr Biochem* 2009;20:743-52.
27. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature* 2009;457:480-4.

28. Turnbaugh PJ, Ridaura VK, Faith JJ, et al. The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. *Science Translational Medicine* 2009;1:6ra14-6ra14.
29. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. *J Physiol* 2009;587:4153-8.
30. Ley RE. Obesity and the human microbiome. *Curr Opin Gastroenterol* 2010;26:5-11.
31. Hildebrandt MA, Hoffmann C, Sherrill-Mix SA, et al. High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* 2009;137:1716-24 e1-2.
32. Zeisel SH, da Costa KA. Choline: an essential nutrient for public health. *Nutr Rev* 2009;67:615-23.
33. Bidulescu A, Chambless LE, Siega-Riz AM, et al. Repeatability and measurement error in the assessment of choline and betaine dietary intake: the Atherosclerosis Risk in Communities (ARIC) study. *Nutr J* 2009;8:14.
34. Xu X, Gammon MD, Zeisel SH, et al. High intakes of choline and betaine reduce breast cancer mortality in a population-based study. *FASEB J* 2009;23:4022-8.
35. Zeisel SH. Genetic polymorphisms in methyl-group metabolism and epigenetics: lessons from humans and mouse models. *Brain Res* 2008;1237:5-11.
36. Song J, da Costa KA, Fischer LM, et al. Polymorphism of the PEMT gene and susceptibility to nonalcoholic fatty liver disease (NAFLD). *FASEB J* 2005;19:1266-71.
37. Zeisel SH, da Costa KA, Fox JG. Endogenous formation of dimethylamine. *Biochem. J.* 1985;232:403-408.
38. Zeisel SH, daCosta KA, LaMont JT. Mono-, di- and trimethylamine in human gastric fluid: potential substrates for nitrosodimethylamine formation. *Carcinogenesis* 1988;9:179-81.
39. Zeisel SH, daCosta KA, Youssef M, et al. Conversion of dietary choline to trimethylamine and dimethylamine in rats: dose-response relationship. *J. Nutr.* 1989;119:800-4.
40. Sohlenkamp C, Lopez-Lara IM, Geiger O. Biosynthesis of phosphatidylcholine in bacteria. *Prog Lipid Res* 2003;42:115-62.
41. Martinez-Morales F, Schobert M, Lopez-Lara IM, et al. Pathways for phosphatidylcholine biosynthesis in bacteria. *Microbiology* 2003;149:3461-71.

42. Aktas M, Wessel M, Hacker S, et al. Phosphatidylcholine biosynthesis and its significance in bacteria interacting with eukaryotic cells. *Eur J Cell Biol* 2010.
43. Sha W, da Costa KA, Fischer LM, et al. Metabolomic profiling can predict which humans will develop liver dysfunction when deprived of dietary choline. *FASEB J* 2010.
44. Backhed F, Crawford PA. Coordinated regulation of the metabolome and lipidome at the host-microbial interface. *Biochim Biophys Acta* 2009.
45. Wikoff WR, Anfora AT, Liu J, et al. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc Natl Acad Sci U S A* 2009;106:3698-703.
46. Claus SP, Tsang TM, Wang Y, et al. Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes. *Mol Syst Biol* 2008;4:219.
47. Cani PD, Bibiloni R, Knauf C, et al. Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. *Diabetes* 2008;57:1470-81.
48. Yap IK, Li JV, Saric J, et al. Metabonomic and microbiological analysis of the dynamic effect of vancomycin-induced gut microbiota modification in the mouse. *J Proteome Res* 2008;7:3718-28.
49. Busby MG, Fischer L, da Costa KA, et al. Choline- and betaine-defined diets for use in clinical research and for the management of trimethylaminuria. *J Am Diet Assoc* 2004;104:1836-45.
50. Dietary intakes for folate, thiamin, riboflavin, niacin, vitamin B12, panthothenic acid, biotin, and choline. In: Institute of Medicine NAPU, ed. Washington, DC: National Academy Press, 1998:390-422.
51. Zeisel SH, Mar MH, Howe JC, et al. Concentrations of choline-containing compounds and betaine in common foods. *J Nutr* 2003;133:1302-7.
52. Koc H, Mar MH, Ranasinghe A, et al. Quantitation of choline and its metabolites in tissues and foods by liquid chromatography/electrospray ionization-isotope dilution mass spectrometry. *Anal Chem* 2002;74:4734-40.
53. da Costa KA, Badea M, Fischer LM, et al. Elevated serum creatine phosphokinase in choline-deficient humans: mechanistic studies in C2C12 mouse myoblasts. *Am J Clin Nutr* 2004;80:163-70.
54. da Costa KA, Gaffney CE, Fischer LM, et al. Choline deficiency in mice and humans is associated with increased plasma homocysteine concentration after a methionine load. *Am J Clin Nutr* 2005;81:440-4.

55. Fishbein MH, Gardner KG, Potter CJ, et al. Introduction of fast MR imaging in the assessment of hepatic steatosis. *Magn Reson Imaging* 1997;15:287-93.
56. Fisher MM, Triplett EW. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* 1999;65:4630-6.
57. Fierer N, Hamady M, Lauber CL, et al. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci U S A* 2008;105:17994-9.
58. Cole JR, Chai B, Farris RJ, et al. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucl. Acids Res.* 2007;35:D169-D172.
59. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;73:5261-7.
60. Liu Z, DeSantis TZ, Andersen GL, et al. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 2008;36:e120.
61. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. *Science* 2005;308:1635-8.
62. Kunin V, Engelbrektsen A, Ochman H, et al. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 2010;12:118-23.
63. Lozupone C, Knight R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* 2005;71:8228-8235.
64. Selan L, Palma S, Scoarughi GL, et al. Phosphorylcholine impairs susceptibility to biofilm formation of hydrogel contact lenses. *Am J Ophthalmol* 2009;147:134-9.
65. Browning JD, Szczepaniak LS, Dobbins R, et al. Prevalence of hepatic steatosis in an urban population in the United States: Impact of ethnicity. *Hepatology* 2004;40:1387-1395.
66. Wieckowska A, Feldstein AE. Diagnosis of nonalcoholic fatty liver disease: invasive versus noninvasive. *Semin Liver Dis* 2008;28:386-95.
67. Zeisel SH. Gene response elements, genetic polymorphisms and epigenetics influence the human dietary requirement for choline. *IUBMB Life* 2007;59:380-7.

68. da Costa KA, Kozyreva OG, Song J, et al. Common genetic polymorphisms affect the human requirement for the nutrient choline. *FASEB J* 2006;20:1336-44.
69. Raetz CR, Whitfield C. Lipopolysaccharide endotoxins. *Annu Rev Biochem* 2002;71:635-700.
70. Kudo H, Takahara T, Yata Y, et al. Lipopolysaccharide triggered TNF-alpha-induced hepatocyte apoptosis in a murine non-alcoholic steatohepatitis model. *J Hepatol* 2009;51:168-75.
71. Higuchi R, Dollinger G, Walsh PS, et al. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y)* 1992;10:413-7.
72. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5:113.
73. Maheux AF, Picard FJ, Boissinot M, et al. Analytical comparison of nine PCR primer sets designed to detect the presence of *Escherichia coli*/*Shigella* in water samples. *Water Res* 2009;43:3019-28.
74. Sundaram SS, Zeitler P, Nadeau K. The metabolic syndrome and nonalcoholic fatty liver disease in children. *Curr Opin Pediatr* 2009;21:529-35.
75. Siegel AB, Zhu AX. Metabolic syndrome and hepatocellular carcinoma: two growing epidemics with a potential link. *Cancer* 2009;115:5651-61.
76. Leite NC, Salles GF, Araujo AL, et al. Prevalence and associated factors of non-alcoholic fatty liver disease in patients with type-2 diabetes mellitus. *Liver Int* 2009;29:113-9.
77. Stickel F, Hellerbrand C. Non-alcoholic fatty liver disease as a risk factor for hepatocellular carcinoma: mechanisms and implications. *Gut* 2010;59:1303-7.
78. Xu X, Gammon MD, Zhang Y, et al. BRCA1 promoter methylation is associated with increased mortality among women with breast cancer. *Breast Cancer Res Treat* 2009;115:397-404.
79. Targher G, Bertolini L, Padovani R, et al. Prevalence of non-alcoholic fatty liver disease and its association with cardiovascular disease in patients with type 1 diabetes. *J Hepatol* 2010;53:713-8.
80. Armougom F, Bittar F, Stremmer N, et al. Microbial diversity in the sputum of a cystic fibrosis patient studied with 16S rDNA pyrosequencing. *Eur J Clin Microbiol Infect Dis* 2009;28:1151-4.
81. Guss AM, Roeselers G, Newton IL, et al. Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. *ISME J* 2011;5:20-9.

82. van der Gast CJ, Walker AW, Stressmann FA, et al. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME J* 2010.
83. About Cystic Fibrosis: What You Need to Know. Volume 2010: Cystic Fibrosis Foundation, 2010.
84. Gibson RL, Burns JL, Ramsey BW. Pathophysiology and management of pulmonary infections in cystic fibrosis. *Am J Respir Crit Care Med* 2003;168:918-51.
85. Rubin BK. Mucus, phlegm, and sputum in cystic fibrosis. *Respir Care* 2009;54:726-32; discussion 732.
86. Smith JJ, Travis SM, Greenberg EP, et al. Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. *Cell* 1996;85:229-36.
87. Gomez MI, Prince A. Opportunistic infections in lung disease: Pseudomonas infections in cystic fibrosis. *Curr Opin Pharmacol* 2007;7:244-51.
88. Lyczak JB, Cannon CL, Pier GB. Lung infections associated with cystic fibrosis. *Clin Microbiol Rev* 2002;15:194-222.
89. Chmiel JF, Davis PB. State of the art: why do the lungs of patients with cystic fibrosis become infected and why can't they clear the infection? *Respir Res* 2003;4:8.
90. Tunney MM, Field TR, Moriarty TF, et al. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. *Am J Respir Crit Care Med* 2008;177:995-1001.
91. Harris JK, De Groote MA, Sagel SD, et al. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. *Proc Natl Acad Sci U S A* 2007;104:20529-33.
92. Rogers GB, Skelton S, Serisier DJ, et al. Determining cystic fibrosis-affected lung microbiology: comparison of spontaneous and serially induced sputum samples by use of terminal restriction fragment length polymorphism profiling. *J Clin Microbiol* 2010;48:78-86.
93. Klepac-Ceraj V, Lemon KP, Martin TR, et al. Relationship between cystic fibrosis respiratory tract bacterial communities and age, genotype, antibiotics and *Pseudomonas aeruginosa*. *Environ Microbiol* 2010.
94. Hoiby N. *Pseudomonas aeruginosa* infection in cystic fibrosis. Diagnostic and prognostic significance of pseudomonas aeruginosa precipitins determined by means of crossed immunoelectrophoresis. A survey. *Acta Pathol Microbiol Scand Suppl* 1977:1-96.

95. Smith EE, Buckley DG, Wu Z, et al. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A* 2006;103:8487-92.
96. Rajan S, Saiman L. Pulmonary infections in patients with cystic fibrosis. *Semin Respir Infect* 2002;17:47-56.
97. Muhdi K, Edenborough FP, Gumery L, et al. Outcome for patients colonised with *Burkholderia cepacia* in a Birmingham adult cystic fibrosis clinic and the end of an epidemic. *Thorax* 1996;51:374-7.
98. Rogers GB, Carroll MP, Bruce KD. Studying bacterial infections through culture-independent approaches. *J Med Microbiol* 2009;58:1401-18.
99. Rogers GB, Carroll MP, Serisier DJ, et al. Characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16s ribosomal DNA terminal restriction fragment length polymorphism profiling. *J Clin Microbiol* 2004;42:5176-83.
100. Rogers GB, Hart CA, Mason JR, et al. Bacterial diversity in cases of lung infection in cystic fibrosis patients: 16S ribosomal DNA (rDNA) length heterogeneity PCR and 16S rDNA terminal restriction fragment length polymorphism profiling. *J Clin Microbiol* 2003;41:3548-58.
101. Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. *Science* 2006;312:1355-9.
102. Rogers GB, Hoffman LR, Whiteley M, et al. Revealing the dynamics of polymicrobial infections: implications for antibiotic therapy. *Trends Microbiol* 2010;18:357-64.
103. Rogers GB, Stressmann FA, Walker AW, et al. Lung infections in cystic fibrosis: deriving clinical insight from microbial complexity. *Expert Rev Mol Diagn* 2010;10:187-96.
104. Duan K, Dammel C, Stein J, et al. Modulation of *Pseudomonas aeruginosa* gene expression by host microflora through interspecies communication. *Mol Microbiol* 2003;50:1477-91.
105. Dunne C. Adaptation of bacteria to the intestinal niche: probiotics and gut disorder. *Inflamm Bowel Dis* 2001;7:136-45.
106. Grice EA, Kong HH, Conlan S, et al. Topographical and temporal diversity of the human skin microbiome. *Science* 2009;324:1190-2.
107. Stecher B, Chaffron S, Kappeli R, et al. Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathog* 2010;6:e1000711.

108. McKone EF, Goss CH, Aitken ML. CFTR genotype as a predictor of prognosis in cystic fibrosis. *Chest* 2006;130:1441-7.
109. Dorfman R, Sandford A, Taylor C, et al. Complex two-gene modulation of lung disease severity in children with cystic fibrosis. *J Clin Invest* 2008;118:1040-9.
110. Rogers GB, Carroll MP, Serisier DJ, et al. Bacterial activity in cystic fibrosis lung infections. *Respir Res* 2005;6:49.
111. Sibley CD, Parkins MD, Rabin HR, et al. A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients. *Proc Natl Acad Sci U S A* 2008;105:15070-5.
112. Hamady M, Walker JJ, Harris JK, et al. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Meth* 2008;5:235-237.
113. Hamp TJ, Jones WJ, Fodor AA. The effects of experimental choices and analysis noise on surveys of the rare biosphere. *Appl Environ Microbiol* 2009.
114. Huse SM, Welch DM, Morrison HG, et al. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 2010.
115. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;215:403-10.
116. Ye Y. AbundantOTU. Volume 2010, 2010: Consensus alignment algorithm for analysis of 16S rRNA pyrosequences.
117. Microbiome Utilities Portal of the Broad Institute. Volume 2010, 2010.
118. Hamady M, Lozupone C, Knight R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 2010;4:17-27.
119. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;23:2947-8.
120. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;35:7188-96.
121. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 2009;10:356.
122. Pai VB, Nahata MC. Efficacy and safety of aerosolized tobramycin in cystic fibrosis. *Pediatr Pulmonol* 2001;32:314-27.

123. Chuchalin A, Amelina E, Bianco F. Tobramycin for inhalation in cystic fibrosis: Beyond respiratory improvements. *Pulm Pharmacol Ther* 2009;22:526-32.
124. Oberhardt MA, Goldberg JB, Hogardt M, et al. Metabolic network analysis of *Pseudomonas aeruginosa* during chronic cystic fibrosis lung infection. *J Bacteriol* 2010;192:5534-48.
125. Doring G, Conway SP, Heijerman HG, et al. Antibiotic therapy against *Pseudomonas aeruginosa* in cystic fibrosis: a European consensus. *Eur Respir J* 2000;16:749-67.
126. Dales L, Ferris W, Vandemheen K, et al. Combination antibiotic susceptibility of biofilm-grown *Burkholderia cepacia* and *Pseudomonas aeruginosa* isolated from patients with pulmonary exacerbations of cystic fibrosis. *Eur J Clin Microbiol Infect Dis* 2009;28:1275-9.
127. Govan JR, Brown AR, Jones AM. Evolving epidemiology of *Pseudomonas aeruginosa* and the *Burkholderia cepacia* complex in cystic fibrosis lung infection. *Future Microbiol* 2007;2:153-64.
128. Worlitzsch D, Rintelen C, Bohm K, et al. Antibiotic-resistant obligate anaerobes during exacerbations of cystic fibrosis patients. *Clin Microbiol Infect* 2009;15:454-60.
129. Martin DW, Schurr MJ, Mudd MH, et al. Mechanism of conversion to mucoidy in *Pseudomonas aeruginosa* infecting cystic fibrosis patients. *Proc Natl Acad Sci U S A* 1993;90:8377-81.
130. Luzar MA, Montie TC. Avirulence and altered physiological properties of cystic fibrosis strains of *Pseudomonas aeruginosa*. *Infect Immun* 1985;50:572-6.
131. Oliver A, Canton R, Campo P, et al. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* 2000;288:1251-4.
132. Mena A, Smith EE, Burns JL, et al. Genetic adaptation of *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients is catalyzed by hypermutation. *J Bacteriol* 2008;190:7910-7.
133. Hoiby N, Bjarnsholt T, Givskov M, et al. Antibiotic resistance of bacterial biofilms. *Int J Antimicrob Agents* 2010;35:322-32.
134. Hassett DJ, Korfhagen TR, Irvin RT, et al. *Pseudomonas aeruginosa* biofilm infections in cystic fibrosis: insights into pathogenic processes and treatment strategies. *Expert Opin Ther Targets* 2010;14:117-30.
135. King P, Citron DM, Griffith DC, et al. Effect of oxygen limitation on the in vitro activity of levofloxacin and other antibiotics administered by the aerosol route

- against *Pseudomonas aeruginosa* from cystic fibrosis patients. *Diagn Microbiol Infect Dis* 2010;66:181-6.
136. Hassett DJ, Sutton MD, Schurr MJ, et al. *Pseudomonas aeruginosa* hypoxic or anaerobic biofilm infections within cystic fibrosis airways. *Trends Microbiol* 2009;17:130-8.
  137. Bittar F, Rolain JM. Detection and accurate identification of new or emerging bacteria in cystic fibrosis patients. *Clin Microbiol Infect* 2010;16:809-20.
  138. Ellis CN, Cooper VS. Experimental adaptation of *Burkholderia cenocepacia* to onion medium reduces host range. *Appl Environ Microbiol* 2010;76:2387-96.
  139. Hoiby N. Hemophilus influenzae, Staphylococcus aureus, *Pseudomonas cepacia*, and *Pseudomonas aeruginosa* in patients with cystic fibrosis. *Chest* 1988;94:97S-103S.
  140. Hoiby N. Isolation and treatment of cystic fibrosis patients with lung infections caused by *Pseudomonas* (*Burkholderia*) *cepacia* and multiresistant *Pseudomonas aeruginosa*. *Neth J Med* 1995;46:280-7.
  141. Baldwin A, Sokol PA, Parkhill J, et al. The *Burkholderia cepacia* epidemic strain marker is part of a novel genomic island encoding both virulence and metabolism-associated genes in *Burkholderia cenocepacia*. *Infect Immun* 2004;72:1537-47.
  142. Mahenthiralingam E, Vandamme P. Taxonomy and pathogenesis of the *Burkholderia cepacia* complex. *Chron Respir Dis* 2005;2:209-17.
  143. Sibley CD, Duan K, Fischer C, et al. Discerning the complexity of community interactions using a *Drosophila* model of polymicrobial infections. *PLoS Pathog* 2008;4:e1000184.
  144. Ulrich M, Beer I, Braitmaier P, et al. Relative contribution of *Prevotella intermedia* and *Pseudomonas aeruginosa* to lung pathology in airways of patients with cystic fibrosis. *Thorax* 2010.
  145. Bittar F, Richet H, Dubus JC, et al. Molecular detection of multiple emerging pathogens in sputa from cystic fibrosis patients. *PLoS One* 2008;3:e2908.
  146. Sibley CD, Parkins MD, Rabin HR, et al. The relevance of the polymicrobial nature of airway infection in the acute and chronic management of patients with cystic fibrosis. *Curr Opin Investig Drugs* 2009;10:787-94.
  147. Benedict KF, Mac Gabhann F, Amanfu RK, et al. Systems Analysis of Small Signaling Modules Relevant to Eight Human Diseases. *Ann Biomed Eng* 2010.

148. Spasenovski T, Carroll MP, Payne MS, et al. Molecular analysis of diversity within the genus *Pseudomonas* in the lungs of cystic fibrosis patients. *Diagn Microbiol Infect Dis* 2009;63:261-7.
149. Knapp BA, Seeber J, Rief A, et al. Bacterial community composition of the gut microbiota of *Cylindroiulus fulviceps* (diplopoda) as revealed by molecular fingerprinting and cloning. *Folia Microbiol (Praha)* 2010;55:489-96.
150. Brown MV, Schwalbach MS, Hewson I, et al. Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. Volume 7, 2005:1466-1479.
151. Dahllöf I. Molecular community analysis of microbial diversity. *Curr Opin Biotechnol* 2002;13:213-7.
152. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;74:5463-7.
153. McCafferty JW, Reid RW, Spencer MD, et al. Peak Studio: a tool for the visualization and analysis of fragment analysis files., 2011.
154. Richterich P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res* 1998;8:251-9.
155. Ewing B, Hillier L, Wendl MC, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998;8:175-85.
156. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998;8:186-94.
157. DeSantis TZ, Jr., Hugenholtz P, Keller K, et al. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* 2006;34:W394-9.
158. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;72:5069-72.
159. Cole JR, Wang Q, Cardenas E, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009;37:D141-5.
160. Khaw KT, Wareham N, Bingham S, et al. Combined impact of health behaviours and mortality in men and women: the EPIC-Norfolk prospective population study. *PLoS Med* 2008;5:e12.

161. Danaei G, Ding EL, Mozaffarian D, et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS Med* 2009;6:e1000058.
162. Ford ES, Mokdad AH. Epidemiology of obesity in the Western Hemisphere. *J Clin Endocrinol Metab* 2008;93:S1-8.
163. Lightfoot JT, Turner MJ, Pomp D, et al. Quantitative trait loci for physical activity traits in mice. *Physiol Genomics* 2008;32:401-8.
164. Lightfoot JT, Turner MJ, Daves M, et al. Genetic influence on daily wheel running activity level. *Physiol Genomics* 2004;19:270-6.
165. Lightfoot JT, Leamy L, Pomp D, et al. Strain screen and haplotype association mapping of wheel running in inbred mouse strains. *J Appl Physiol* 2010;109:623-34.
166. Li M, Wang B, Zhang M, et al. Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A* 2008;105:2117-22.
167. Cani PD, Delzenne NM. Gut microflora as a target for energy and metabolic homeostasis. *Curr Opin Clin Nutr Metab Care* 2007;10:729-34.
168. Hoffman-Goetz L, Pervaiz N, Guan J. Voluntary exercise training in mice increases the expression of antioxidant enzymes and decreases the expression of TNF-alpha in intestinal lymphocytes. *Brain Behav Immun* 2009;23:498-506.
169. Hoffman-Goetz L, Pervaiz N, Packer N, et al. Freewheel training decreases pro- and increases anti-inflammatory cytokine expression in mouse intestinal lymphocytes. *Brain Behav Immun* 2010;24:1105-15.
170. Hagio M, Matsumoto M, Yajima T, et al. Voluntary wheel running exercise and dietary lactose concomitantly reduce proportion of secondary bile acids in rat feces. *J Appl Physiol* 2010;109:663-8.
171. Reid RW SM, Hamp TJ, Fodor AA. Clusters generated by ARISA data are robust to choices made during binning., 2011.
172. Leamy LJ, Pomp D, Lightfoot JT. An epistatic genetic basis for physical activity traits in mice. *J Hered* 2008;99:639-46.
173. Leamy LJ, Pomp D, Lightfoot JT. Genetic variation in the pleiotropic association between physical activity and body weight in mice. *Genet Sel Evol* 2009;41:41.
174. Leamy LJ, Pomp D, Lightfoot JT. Genetic variation for body weight change in mice in response to physical exercise. *BMC Genet* 2009;10:58.

175. Leamy LJ, Pomp D, Lightfoot JT. A search for quantitative trait loci controlling within-individual variation of physical activity traits in mice. *BMC Genet* 2010;11:83.
176. Ley RE, Backhed F, Turnbaugh P, et al. Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* 2005;102:11070-5.
177. Matsumoto M, Inoue R, Tsukahara T, et al. Voluntary running exercise alters microbiota composition and increases n-butyrate concentration in the rat cecum. *Biosci Biotechnol Biochem* 2008;72:572-6.
178. Sarma-Rupavtarm RB, Ge Z, Schauer DB, et al. Spatial distribution and stability of the eight microbial species of the altered schaedler flora in the mouse gastrointestinal tract. *Appl Environ Microbiol* 2004;70:2791-800.

## APPENDIX A: SUPPLEMENTAL MATERIALS AND METHODS FOR CHAPTER 1

### *Study Diets*

Study diets were composed of 0.8 g/kg high biologic value protein (current Dietary Reference Intake [DRI]), 30% of energy from fat, and the remaining energy from carbohydrate. Total food intake was adjusted to be isocaloric and to provide adequate intakes of macro- and micronutrients (supplemental table 2). All diets met or exceeded the Estimated Average Requirement for methionine plus cysteine and the DRI for vitamins B-12, B-6, and folic acid. A multivitamin supplement (Kirkland Brand Multivitamin Multi-mineral) provided vitamins A, D, B-12, and C and thiamine, niacin, and riboflavin at or above the DRI. Subjects were also given a calcium and magnesium supplement. Subjects' weights remained relatively stable, as caloric intake was adjusted for each subject over the course of the study to maintain subject weight (supplemental table 9). In addition to changes in choline levels, all depletion diets included a soy shake containing Benefiber and, thereby, differed from the study diets at other time points. Description of and micronutrient levels of all diets are extensively detailed in Busby et al.<sup>1</sup>

### *ARISA Analysis Methods*

Using extracted DNA, the intergenic region between the 16S rRNA and the 23S rRNA genes was PCR amplified for each of two technical replicates using universal bacterial primers 1406F-FAM (FAM+TGY ACA CAC CGC CCG T) and 125R (GGG TTB CCC CAT TCR G). Reactions were set up using 50ng of template DNA according

to a NanoDrop ND-1000 spectrophotometer. Thermalcycling conditions were as follows: an initial denaturation step at 94°C for 2 minutes, followed by 35 cycles of 94°C for 25 seconds; 56.5C for 30 seconds; 72°C for 60 seconds. An extension was carried out at 72°C for 5 minutes. Samples were loaded on an Applied Biosystems 3130 genetic analyzer. Applied Biosystems GeneScan™ 1200 LIZ® size standard was used to determine sizing up to 1200 nucleotides. The labeled fragments were separated by size on an Applied Biosystems 3130 sequencer, resulting in an electropherogram where each signal peak represented a different bacterial signature (supplemental figure 1). The electropherogram from each replicate was analyzed using custom JAVA software that establishes a base-pair scale, calls the data peaks from the spectra, assigns base pair sizes to each peak and provides summaries of the peak call assignments. Peak calls were validated using Applied Biosystems 3130 analysis software. Resulting data vectors from all samples were compared using hierarchical clustering with custom JAVA code.

### ***Sequencing Analysis Methods***

The V1-V2 variable regions of the 16S rRNA gene were targeted using the 454 Life Sciences primer B with a “TC” linker and bacterial 27F primer (5'-GCCTTGCCAGCCCGCTCAGTCAGAGTTTGATCCTGGCTCAG-3') and 454 Life Sciences primer A with a “CA” linker, 12 mer barcode and bacterial primer 338R (5'-GCCTCCCTCGCGCCATCAGNNNNNNNNNNNNNCATGCTGCCTCCCGTAGGAGT-3') where the N's represent barcodes used to identify each sample.<sup>2</sup> PCRs were set up with Platinum Taq DNA polymerase (Invitrogen) according to the included protocol with 100ng of bacterial genomic DNA as a template. Each reaction was quantitated by

PicoGreen on a NanoDrop ND-3300 fluorospectrometer. Samples were pooled in equimolar amounts and concentrated in a vacuum centrifuge before being submitted to the Environmental Genomics Core Facility at the University of South Carolina for 454-FLX sequencing.

Over 200,000 sequences were obtained from the 454-FLX process and were end-trimmed based on the Lucy algorithm at a cut-off of 0.002 corresponding to a quality score of 27.<sup>3</sup> Sequences had to meet the following criteria<sup>4</sup> to be included in the final dataset: (1) no Ns in the trimmed sequence, (2) an exact match to the 5' primer, (3) Lucy's identified region of poor quality at the 0.002 threshold did not extend beyond the 5' primer. The 5' primer (including the barcode) was trimmed from the sequences before analysis. Any sequences that did not meet a length requirement from 180 to 280 bases after trimming were discarded. The 194,781 trimmed, quality-controlled sequences (supplemental table 4) were evaluated for human contamination by using BLASTn<sup>5</sup> searches against the *E. coli* (J01695) 16S rRNA gene. All but 60 of our sequences matched the *E. coli* 16S rRNA gene with an e-score threshold of 0.04. For the 60 sequences that did not meet this threshold, a BLASTn search was performed against the entire bacterial 16S rRNA Ribosomal Database Project (RDP)<sup>6</sup> database. Every one of these sequenced matched this database with an e-value of 2.00E-90 or smaller. Taken together, these results indicate that our quality-controlled, trimmed sequences had little to no human contamination. Rarefaction curves illustrate that the resulting sequences provided sufficient sequence depth in all subjects and in 74 of the 79 samples originally collected (supplemental figures 2-3).

### ***Statistical Methods***

The standardized logged sequence abundance was used in statistical analysis of all sequence counts, using the following calculation:

$$\text{LOG}_{10} ((\text{Frequency} / \# \text{ sequences in sample}) * \text{Average} \# \text{ of sequences per sample} + 1)$$

As an example, consider a dataset in which there was an average of 2,500 sequences per sample assigned to phylum at a 50% confidence level. For one sample within this dataset, 1,000 sequences were assigned to phylum and, of those, 300 were assigned to the taxon Firmicutes. The transformation would be:

$$\text{Log}_{10} (( 300 / 1000 ) * 2500 + 1 ) = 2.8756$$

Using this measure, the logged sequence proportion corrects for different samples having different total numbers of sequences.

For hierarchical clustering of OTUs, we estimated the probability of obtaining the perfect clustering of samples by subject that we observed by sequencing by calculating how many different ways there are to achieve perfect clustering and dividing this value by the total number of ways the samples could be clustered. For the purposes of this calculation, we consider the hierarchical cluster as arranging the samples in an unweighted tree and the results of the cluster as an ordered set of our 74 samples represented by the 74 most derived nodes of the tree. In a tree with perfect clustering by subject, there are  $x_i!$  possible arrangements of samples for each subject, where  $x_i$  is the number of samples for subject  $i$  (supplemental table 4). Since there are 15 ways of arranging our subjects while maintaining perfect clusters, the total number of unweighted trees that we could have observed that would have contained perfect clustering by patient is given by:

$$15! \prod_{i=1}^{i=15} X_i \quad (1)$$

Since there are 74 samples, there are 74! possible arrangements of all of our samples. Dividing equation (1) by 74! yields a value of  $6.78e^{-65}$ . We conclude that we can reject a null hypothesis that we could have observed perfect clustering of our patients by chance with a p-value of essentially zero. For all taxonomic statistical testing, we report results at the most inclusive taxonomic level for which effects were identified. From a practical perspective, we were able to determine that statistical testing at fine levels of detail was unlikely to be productive. In many cases, a particular family or genus would only be present in a small subset of our subjects, sometimes only in one subject, which reduced sample size and, therefore, the power of any statistical analysis, to the point where achieving significance was improbable. We were also aware of the problem created by multiple comparisons and wanted to approach our dataset with that issue in mind. To manage power and to provide some control for multiple comparisons, we chose to conduct statistical testing at the highest taxonomic level present in at least half of our subjects. While we recognized that the range of bacterial types within a group as large as class would likely also represent a range of functional types, our observed correlations between physiological changes and bacterial abundance suggested that that function in the bacterial group was consistent enough to support the host physiology link.

For paired t-tests comparing time points, samples missing one or more time points (subjects 29, 04) were excluded from the analysis. To manage the number of comparisons for which we had to correct, we removed from statistical analysis any taxon which did not have a presence in at least 50% of our subjects (9 categories for the

taxonomic class level), a criterion which reduced the possibility of spurious observations that might result from low sample sizes.

To correct for multiple comparisons in statistical testing, we used the Benjamini and Hochberg algorithm<sup>7</sup> to adjust each p-value. The adjusted p-values for a statistical test run over a set of taxa (ordered in ascending rank by p-value) is given by  $N * p(k)/k$  where N is the number of taxa for which a null hypothesis is evaluated by the test statistic, p(k) is the p-value produced by the test statistic and k is the rank of the taxa within the p-value ranked list. (Note that, for the top hit with the smallest p-value, the adjusted p-value is identical to Bonferroni correction). The adjusted p-value can then be evaluated against a threshold false discovery rate. For example, an adjusted p-value of 0.05 indicates that a result could be considered significant if a commonly-selected 5% FDR threshold was chosen. We note that our adjusted p-values can be greater than 1 and hence cannot be strictly considered as classical p-values.

In our PCA models of the relationships between bacterial abundance, host genotype and liver fat changes, we recognized the potential problem of model over-fitting. Because we selected the two taxa having the highest  $R^2$ s of the nine regressions we performed (one for each class found in at least 8 of our patients), we would expect that the first component of the PCA could be well matched to the LF:SF ratio % change. We recognize that the reported p-value from the linear regression in Figure 5C is, therefore, likely to be anticonservative. To correct for model over-fitting, we developed a permutation procedure which randomly reassigned the taxa associated with each subject, then performed regressions between the permuted taxa and the LF:SF % change. (For example, in a permutation the taxa that were associated with subject 04, could be

assigned to patient 28; for 15 subjects, the number of possible permutations for this procedure is  $15!$  or  $1.31\text{E}+12$ .) We then selected the two taxa that had regressions with the highest  $R^2$  values for each permutation and used those taxa in a PCA. We then computed the  $R^2$  of the first component of that PCA against the LF:SF ratio % change. We ran the permutation procedure one million times. The “permuted p-value” reported from this process is the fraction of times that we observed an  $R^2$  with a value greater than or equal to the  $R^2$  we observed in the unpermuted data. To correct for over-fitting in our model in Figure 5D, we performed a similar set of one million permutations to those described for Figure 5C. We maintained the correct SNP genotype assignment for each subject while reassigning taxa to different subjects, as before. The reported permutation based p-value, therefore, tests the null hypothesis that, given the established relationship between the *PEMT* SNP and fatty liver, inclusion of microbiome composition adds no power to the model. All permutations were conducted using custom JAVA software (source code available on request).

**Supplemental References**

1. Busby MG, Fischer L, da Costa KA, et al. Choline- and betaine-defined diets for use in clinical research and for the management of trimethylaminuria. *J Am Diet Assoc* 2004;104:1836-45.
2. Hamady M, Walker JJ, Harris JK, et al. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Meth* 2008;5:235-237.
3. Kunin V, Engelbrektson A, Ochman H, et al. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 2010;12:118-23.
4. Huse SM, Welch DM, Morrison HG, et al. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 2010.
5. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;215:403-10.
6. Cole JR, Chai B, Farris RJ, et al. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucl. Acids Res.* 2007;35:D169-D172.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 1995;57:289-300.

**Supplemental Tables****Supplemental Table 1**

Status Post/ Pre	Study #	Age	BMI	Race Ethnicity	PEMT	Depletion Status		Start Date (Day 1)	End Date (Final Study)	Total Days in Study
						Depleted Y/N	LF:SF % Change			
Post	28	73	22.2	White	HET	Y	54.4	1/5/08	3/6/08	62
Post	29	61	27.8	White	WT	N	-14.6	1/7/08	3/8/08	62
Post	30	67	30.5	Black	HET	Y	33.1	3/9/08	5/9/08	62
Post	31	58	25.3	White	HET	N	3.1	3/12/08	5/12/08	62
Post	32	56	28.3	White	HO	Y	48.5	3/12/08	5/27/08	77
Post	33	52	28.3	Black	WT	N	-17.6	3/26/08	5/26/08	62
Post	34	61	26.5	White	HET	Y	56.3	5/14/08	7/7/08	55
Post	36	50	29.5	Hispanic	HET	N	-8.8	5/28/08	7/27/08	61
Post	37	78	22.9	Hispanic	HET	Y	38.5	7/16/08	9/29/08	76
Post	38	59	23.3	White	HET	Y	44.2	7/16/08	8/25/08	41
Post	39	61	21.4	White	HET	Y	30.3	8/6/08	10/20/08	76
Pre	03	20	25.4	White	HET	N	23.6	5/4/08	7/4/08	62
Post	41	69	26.6	White	WT	N	18.8	9/3/08	11/3/08	62
Post	42	71	26.9	White	WT	N	2	9/20/08	11/20/08	62
Pre	04	49	30.5	Black	WT	N	-5.2	5/28/08	7/28/08	62

**Table 1. Descriptive information by Subject.** For the *PEMT* SNP (promoter region: 12325817), WT = Wild Type, HET = Heterozygous, HO = Homozygous.

**Supplemental Table 2: Nutrients in Baseline, Depletion and Repletion Diets**

Nutrients*	Breakfast			Lunch		
	Baseline	Depletion	Repletion	Baseline	Depletion	Repletion
Choline (mg)	129.10	13.14	243.69	150.45	10.54	152.60
Energy (kcal)	484.29	474.29	377.72	700.47	657.64	640.33
Pro (g)	7.12	13.79	9.70	16.45	15.16	18.66
Fat (g)	18.64	15.65	11.14	24.05	21.70	27.17
Carb (g)	72.65	72.82	59.35	103.58	104.63	78.70
Folate (mcg)	72.24	20.01	100.30	157.28	4.80	158.08
Vitamin B-12 (mcg)	0.00	0.00	0.64	0.17	0.17	0.24
Met (g)	0.06	0.00	0.06	0.32	0.13	0.37
Cys (g)	0.07	0.00	0.06	0.20	0.03	0.22
Na (mg)	546.49	266.38	530.36	2560.95	191.20	2709.85
Water (g)	29.78	244.72	308.24	215.45	116.45	219.36
Fiber - total dietary (g)	3.25	2.38	1.18	1.62	1.44	1.62
Ca (mg)	29.83	12.61	43.15	187.40	147.95	242.60
P (mg)	293.67	117.47	263.99	389.20	111.20	440.50
K (mg)	283.45	743.51	274.11	275.14	652.16	291.54
Fe (mg)	2.79	1.24	2.54	5.71	0.28	5.72
Zn (mg)	1.05	0.85	0.85	1.81	0.66	2.09
Mg (mg)	80.79	74.90	21.86	35.49	9.20	38.19
Cu (mg)	0.14	0.12	0.12	0.37	0.04	0.37
Mn (mg)	1.30	1.61	0.78	0.57	0.09	0.57
Se (mcg)	18.69	8.68	25.81	36.22	3.23	37.66
Vit C (mg)	0.02	0.02	0.27	0.02	25.44	0.02
Thiamin (mg)	0.37	0.14	0.27	0.49	0.02	0.50
Riboflavin (mg)	0.24	0.07	0.45	0.56	0.11	0.59
Niacin (mg)	2.64	0.21	1.92	5.42	0.24	5.43
Pantothenic acid (mg)	0.36	0.20	0.99	0.83	0.20	0.88
Vit B-6 (mg)	0.05	0.03	0.10	0.06	0.05	0.07
Vit A (IU) (IU)	764.76	382.20	436.15	2237.49	617.10	2333.59
Vit E (mg_ATE)	2.43	1.11	0.76	1.32	1.11	1.32
Vit D (IU)	0.00	0.00	17.27	0.00	2.40	0.00
Cholesterol (mg)	0.00	0.00	211.50	46.16	21.00	55.56
Vitamin K (mcg)	12.32	6.51	3.26	8.39	6.11	8.66
Protein (%)	5.85	11.32	10.30	9.45	8.99	11.78
Fat (%)	34.45	28.90	26.64	31.07	28.96	38.57
Carbohydrate (%)	59.70	59.77	63.06	59.48	62.06	49.65
Alcohol (%)	0.00	0.00	0.00	0.00	0.00	0.00

\*Nutrient values do not include supplements provided to subjects (Supplemental Materials and Methods, Study Diets)

**Supplemental Table 2 (cont.)**

Nutrients*	4PM Snack			Dinner		
	Baseline	Depletion	Repletion	Baseline	Depletion	Repletion
Choline (mg)	95.19	2.09	189.67	206.39	19.85	227.68
Energy (kcal)	207.93	220.54	316.32	599.83	521.09	581.57
Pro (g)	1.18	0.16	2.25	24.48	20.11	17.86
Fat (g)	5.00	5.72	10.00	11.02	7.66	11.12
Carb (g)	39.55	43.40	53.61	100.91	92.66	103.17
Folate (mcg)	0.00	0.00	0.00	257.01	89.19	223.70
Vitamin B-12 (mcg)	0.00	0.00	0.00	0.22	0.10	0.14
Met (g)	0.00	0.00	0.00	0.60	0.27	0.40
Cys (g)	0.00	0.00	0.00	0.30	0.15	0.21
Na (mg)	9.48	9.48	9.48	708.16	358.23	556.15
Water (g)	211.17	211.17	211.17	386.85	163.20	361.06
Fiber - total dietary (g)	0.00	0.00	0.00	6.76	0.69	6.33
Ca (mg)	7.11	7.11	7.11	86.30	22.12	76.91
P (mg)	30.81	30.81	30.81	414.80	115.87	287.79
K (mg)	61.11	6.76	119.85	735.64	646.43	684.49
Fe (mg)	0.05	0.05	0.05	4.49	2.19	3.44
Zn (mg)	0.02	0.02	0.02	2.21	1.12	1.74
Mg (mg)	2.37	2.37	2.37	63.35	25.07	54.16
Cu (mg)	0.01	0.01	0.01	0.26	0.12	0.22
Mn (mg)	0.04	0.04	0.04	1.02	0.73	0.92
Se (mcg)	0.24	0.24	0.24	37.51	19.60	26.21
Vit C (mg)	0.00	0.00	0.00	103.35	0.03	103.35
Thiamin (mg)	0.00	0.00	0.00	0.51	0.26	0.39
Riboflavin (mg)	0.00	0.00	0.00	0.44	0.06	0.33
Niacin (mg)	0.00	0.00	0.00	8.10	3.83	5.67
Pantothenic acid (mg)	0.00	0.00	0.00	1.79	0.73	1.55
Vit B-6 (mg)	0.00	0.00	0.00	0.67	0.25	0.55
Vit A (IU) (IU)	0.00	0.00	0.00	2778.25	1.15	2587.30
Vit E (mg_ATE)	0.00	0.00	0.00	1.33	0.01	0.67
Vit D (IU)	0.00	0.00	0.00	0.00	0.00	0.00
Cholesterol (mg)	0.00	0.00	0.00	41.59	16.69	24.99
Vitamin K (mcg)	0.00	0.00	0.00	175.12	0.16	172.30
Protein (%)	2.27	0.29	2.87	16.30	15.47	12.23
Fat (%)	21.65	22.80	28.72	16.51	13.26	17.13
Carbohydrate (%)	76.08	76.91	68.41	67.19	71.28	70.65
Alcohol (%)	0.00	0.00	0.00	0.00	0.00	0.00

\*Nutrient values do not include supplements provided to subjects (Supplemental Materials and Methods, Study Diets)

**Supplemental Table 2 (cont.)**

<b>Nutrients*</b>	<b>Bedtime Snack</b>			<b>Total</b>		
	Baseline	Depletion	Repletion	Baseline	Depletion	Repletion
Choline (mg)	3.38	3.38	10.28	584.51	49.00	823.92
Energy (kcal)	230.66	150.08	106.68	2223.17	2023.64	2022.61
Pro (g)	1.96	1.96	2.55	51.19	51.18	51.01
Fat (g)	9.69	9.69	0.98	68.39	60.41	60.42
Carb (g)	35.60	14.81	22.18	352.28	328.33	317.02
Folate (mcg)	12.60	12.60	47.88	499.12	126.60	529.95
Vitamin B-12 (mcg)	0.00	0.00	0.00	0.39	0.27	1.03
Met (g)	0.03	0.03	0.05	1.00	0.43	0.87
Cys (g)	0.02	0.02	0.05	0.60	0.20	0.55
Na (mg)	182.91	173.32	487.20	4008.00	998.62	4293.05
Water (g)	216.75	699.83	700.22	1060.00	1435.38	1800.05
Fiber - total dietary (g)	1.26	1.26	0.90	12.89	5.76	10.02
Ca (mg)	13.83	104.72	108.08	324.46	294.50	477.84
P (mg)	46.20	46.20	31.64	1174.68	421.55	1054.72
K (mg)	359.37	357.00	40.88	1714.71	2405.87	1410.87
Fe (mg)	0.88	0.46	1.21	13.91	4.22	12.96
Zn (mg)	0.42	0.31	0.24	5.51	2.96	4.94
Mg (mg)	21.13	18.76	9.80	203.14	130.30	126.39
Cu (mg)	0.13	0.09	0.07	0.91	0.38	0.81
Mn (mg)	0.15	0.12	0.50	3.09	2.60	2.82
Se (mcg)	2.51	2.27	1.62	95.16	34.01	91.54
Vit C (mg)	8.71	8.71	0.00	112.10	34.20	103.65
Thiamin (mg)	0.05	0.05	0.13	1.42	0.46	1.28
Riboflavin (mg)	0.06	0.06	0.17	1.29	0.29	1.55
Niacin (mg)	1.07	1.07	1.47	17.24	5.35	14.49
Pantothenic acid (mg)	0.11	0.11	0.08	3.09	1.24	3.50
Vit B-6 (mg)	0.18	0.18	0.03	0.96	0.51	0.75
Vit A (IU) (IU)	0.00	0.00	0.00	5780.51	1000.45	5357.05
Vit E (mg_ATE)	1.37	1.37	0.00	6.45	3.60	2.76
Vit D (IU)	0.00	0.00	0.00	0.00	2.40	17.27
Cholesterol (mg)	0.00	0.00	0.00	87.75	37.69	292.05
Vitamin K (mcg)	0.00	0.00	0.25	195.83	12.79	184.47
Protein (%)	3.30	5.08	9.46	9.18	9.93	10.12
Fat (%)	36.72	56.52	8.19	27.61	26.37	26.97
Carbohydrate (%)	59.97	38.40	82.35	63.21	63.70	62.90
Alcohol (%)	0.00	0.00	0.00	0.00	0.00	0.00

\*Nutrient values do not include supplements provided to subjects (Supplemental Materials and Methods, Study Diets)

**Supplemental Table 3: MRI Measurement Values for Liver Fat and Spleen Fat**

Subject	Study Day	In/out Liver	In/out Spleen	Ratio L/S	% change B1 to D2
		MRI mean	MRI mean		
28	Baseline	107.3	55.9	1.92	
	Depletion	164.2	55.4	2.96	54.4
29	Baseline	97.3	52.3	1.86	
	Depletion	84.2	53.0	1.59	-14.6
30	Baseline	93.7	61.4	1.53	
	Depletion	129.8	63.9	2.03	33.1
31	Baseline	105.1	66.1	1.59	
	Depletion	102.1	62.3	1.64	3.1
32	Baseline	106.8	60.1	1.78	
	Depletion	161.0	61.0	2.64	48.5
33	Baseline	111.6	63.3	1.76	
	Depletion	90.1	62.0	1.45	-17.6
34	Baseline	104.1	79.9	1.30	
	Depletion	164.1	80.6	2.04	56.3
36	Baseline	101.2	75.4	1.34	
	Depletion	97.4	79.6	1.22	-8.8
37	Baseline	100.4	71.2	1.41	
	Depletion	142.6	73.0	1.95	38.5
38	Baseline	120.6	73.4	1.64	
	Depletion	171.3	72.3	2.37	44.2
39	Baseline	98.6	77.0	1.28	
	Depletion	119.8	71.8	1.67	30.3
03	Baseline	102.5	68.1	1.51	
	Depletion	136.2	73.2	1.86	23.6
41	Baseline	145.6	56.1	2.60	

**Supplemental Table 3 (cont.)**

	Depletion	179.2	58.1	3.08	18.8
42	Baseline	100.2	75.8	1.32	
	Depletion	102.9	76.3	1.35	2.0
04	Baseline	94.4	57.7	1.64	
	Depletion	100.8	65.0	1.55	-5.2

**Table 3. Components of the B1 to D2 Liver Fat to Spleen Fat (LF:SF) ratio percent change.** For each subject, MRI measurements of liver fat values, spleen fat values and the ratio of liver fat to spleen fat are provided for Baseline and Depletion time points. The percentage change in the LF:SF ratio from B1 to D2 is also calculated.

**Supplemental Table 4**

<b>Patient #</b>	<b>Usable Samples</b>	<b>Usable Sequences</b>
03	B1, B2, D1, D2, R	8,819
04	B1, B2, D2	7,578
28	B1, B2, D1, D2, R	18,486
29	B1, B2	10,299
30	B1, B2, D1, D2, R	14,448
31	B1, B2, D1, D2, R	6,571
32	B1, B2, D1, D2, R1, R2	14,838
33	B1, B2, D1, D2, R	8,933
34	B1, B2, D1, D2, R1, R2	9,887
36	B1, B2, D1, D2, R	13,321
37	B1, B2, D1, D2, R1, R2	14,970
38	B1, B2, D1, D2, R	13,763
39	B1, B2, D1, D2, R1, R2	20,698
41	B1, B2, D1, D2, R	24,894
42	B1, B2, D1, D2, R	7,366
<b>Total</b>		194,871
<b>Patients</b>		15
<b>Samples</b>		74
<b>Sequences Generated</b>		213,375
<b>Usable Sequences</b>		194,871
<b>OTUs at 97%</b>		4,857

**Table 4. Subject and sample descriptive statistics.**

Usable sequences met quality standards: no Ns in sequence, exact match to 5' primer, exact match to barcode tag, no low quality sequence beyond 5' primer. Sequences that passed quality standards and that had a length 180 nt to 280 nt, after quality trimming, were retained.

**Supplemental Table 5**

<b>Class</b>	<b>R<sup>2</sup></b>	<b>Direction</b>	<b>p-value</b>	<b>adjusted p-value</b>
Gammaproteobacteria	0.5679	negative	0.00118	0.01062
Erysipelotrichi	0.3822	positive	0.01403	0.06314
Deltaproteobacteria	0.1262	positive	0.19383	0.58149
Bacteroidia	0.0992	negative	0.25290	0.56903
Clostridia	0.0253	positive	0.57110	1.02798
Flavobacteria	0.0246	negative	0.57674	0.86511
Betaproteobacteria	0.0062	negative	0.78000	1.00286
Actinobacteria	0.0040	positive	0.82268	0.92552
Bacilli	0.0022	positive	0.86759	0.86759

**Table 5. Results from regressions showing predictive value of B1 bacteria abundance for choline deficiency induced fatty liver development.** Correlations, p-values and adjusted p-values from regressions testing the null hypothesis that the slope of the linear relationship between B1 abundance levels of each class and the percentage change in the liver fat to spleen fat ratio from time point B1 (study initiation) to D2 (maximum subject choline deficiency) is zero. The adjusted p-value (see methods) corrects the p-value for multiple comparisons (n=9) and shows that only Gammaproteobacteria B1 abundance shows a significant linear correlation to liver fat change at a false discovery rate of 0.05.

**Supplemental Table 6**

<b>Unifrac Environmental Distance P-Values</b>						
<b>Patient</b>	<b>B1</b>	<b>B2</b>	<b>D1</b>	<b>D2</b>	<b>R1</b>	<b>R2</b>
28	<=0.01	0.38	1.00	0.99	0.99	***
30	<=0.01	0.99	0.97	.081	1.00	***
31	0.76	1.00	0.96	0.52	<=0.01	***
32	0.60	0.63	0.05	0.59	0.89	0.41
33	<=0.01	0.99	0.74	0.75	0.97	***
34	<=0.01	0.96	1.00	0.81	0.83	0.17
36	0.89	0.05	0.79	0.74	0.97	***
37	0.12	0.93	0.97	0.29	0.09	1.00
38	0.70	0.89	0.58	0.58	0.64	***
39	0.01	0.91	0.98	0.88	0.91	0.17
03	0.06	0.49	0.39	0.57	0.96	***
41	0.02	0.70	0.86	0.71	0.98	***
42	0.27	0.35	0.58	0.39	0.98	***
04	0.83	0.90	***	0.87	***	***

**Table 6. P-values based on Unifrac analysis of samples within subject.**  
The p-value represents the probability that a sample has more unique phylogenetic branch lengths that would be expected by chance. Low values indicate that the sample is different from the other patient samples. Asterisks designate missing samples. Noteworthy are the B1 samples from patients 28, 30, 33 and 34, where baseline samples are distinct.

**Supplemental Table 7**

Classes	B1 to B2		D2 to R		B1 to D2		B1 to R	
	p-value	adjusted p-value	p-value	adjusted p-value	p-value	adjusted p-value	p-value	adjusted p-value
Gammaproteobacteria	0.727	0.793	0.006	0.216	0.523	0.856	0.041	0.492
Betaproteobacteria	0.074	0.533	0.008	0.144	0.221	0.530	0.175	0.525
Bacilli	0.059	0.531	0.627	0.806	0.171	0.560	0.079	0.406
Clostridia	0.078	0.468	0.841	0.841	0.205	0.527	0.124	0.496
Deltaproteobacteria	0.528	0.836	0.665	0.798	0.373	0.707	0.086	0.387
Bacteroidia	0.781	0.803	0.503	0.862	0.151	0.544	0.185	0.512
Erysipelotrichi	0.659	0.818	0.610	0.845	0.239	0.538	0.307	0.650
Actinobacteria	0.326	0.652	0.699	0.812	0.595	0.857	0.440	0.792
Flavobacteria	0.725	0.816	0.551	0.827	0.615	0.820	0.755	0.799

**Table 7. Significance of changes in bacterial frequencies from one time point to another.** P-values and adjusted p-value (see methods) from paired t-tests of the null hypothesis that there was no change between time points in logged standardized sequence abundance for each bacterial class. P-value indicates the probability that the difference could be as significant by chance. Adjusted p-value corrects the p-value for multiple comparisons (n=36) and shows that no difference is significant at a false discovery rate of 0.05.

**Supplemental Table 8: Multivariate Regression Effect Testing**

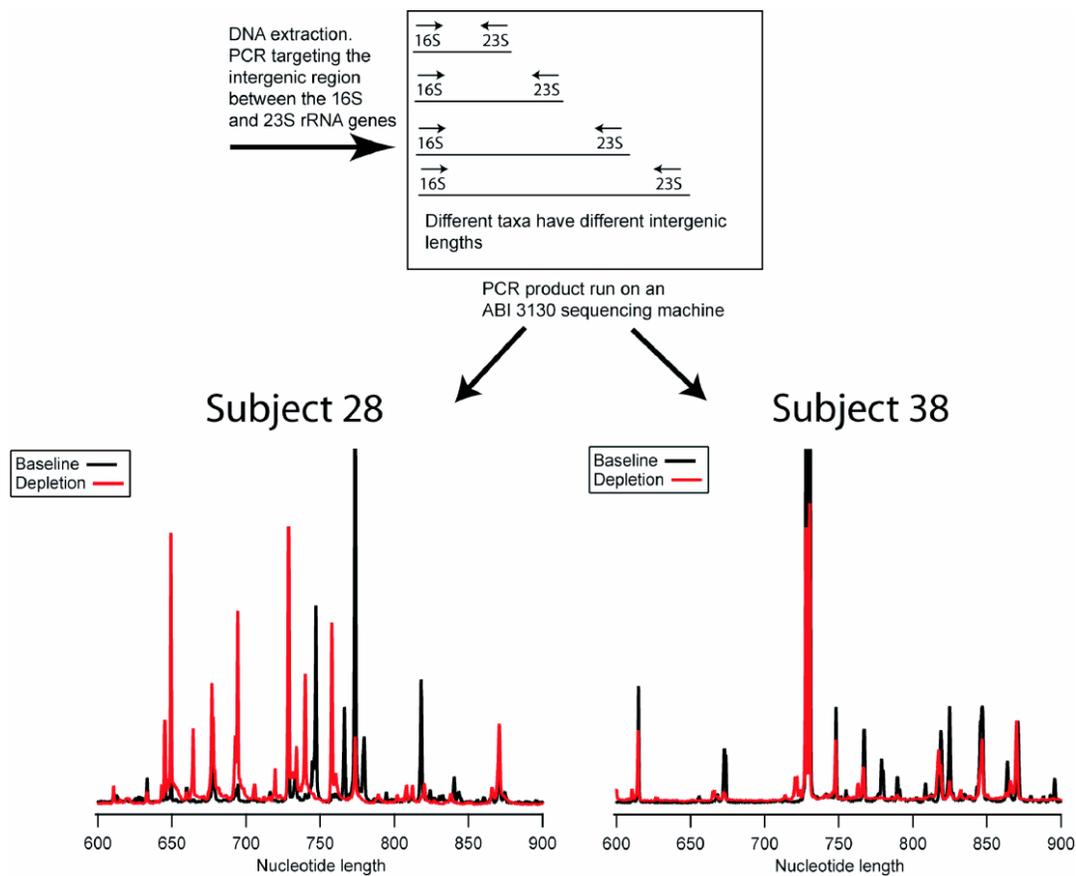
	Sum of Squares	F Ratio	Probability > F
<b>Model 1: R<sup>2</sup>=.6545 P=.0017</b>			
% change LF:SF ratio against:			
Gammaproteobacteria	2509.47	9.46	0.0096
Erysipelotrichi	798.67	3.01	0.1083
<b>Model 2: R<sup>2</sup>=.8716 P=3.3E-5</b>			
% change LF:SF ratio against:			
Gammaproteobacteria	1402.09	13.04	0.0041
Erysipelotrichi	711.16	6.61	0.0260
<i>PEMT</i>	2000.6	18.6	0.0012

**Table 8. Multivariate regression effect testing shows contribution of each factor to model.** Probability > F is the probability that if the null hypothesis is true, a larger F-statistic would occur due to random error – the probability that the actual effect is zero. Model 1 considers only the taxa Gammaproteobacteria and Erysipelotrichi. Model 2 considers both taxa as well as host genotype for *PEMT*

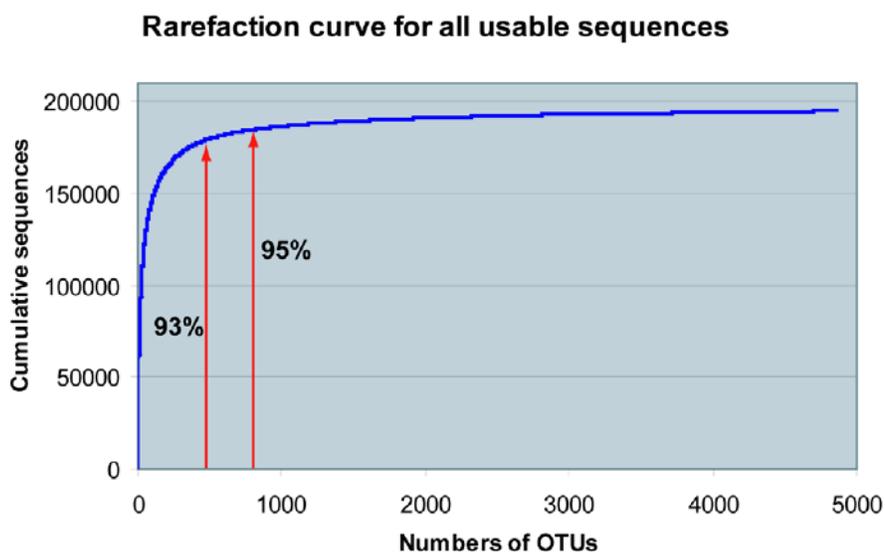
**Supplemental Table 9: Weight of Each Subject at Baseline (B1), Depletion (D2) and Repletion (R) Time Points**

Subject	Baseline (B1) Weight (kg)	Depletion (D2) Weight (kg)	Repletion (R1) Weight (kg)
28	70.6	71.7	72.5
29	77.6	80.6	81.0
30	75.5	76.0	75.8
32	73.3	72.9	71.3
33	72.8	73.8	73.4
34	64.6	65.6	64.6
36	80.9	82.1	83.4
37	53.5	54.4	54.6
38	63.7	62.3	62.5
39	51.0	51.8	52.3
03	71.2	71.5	72.3
41	63.7	64.4	64.8
42	70.9	72.8	71.7
04	84.0	80.8	81.9

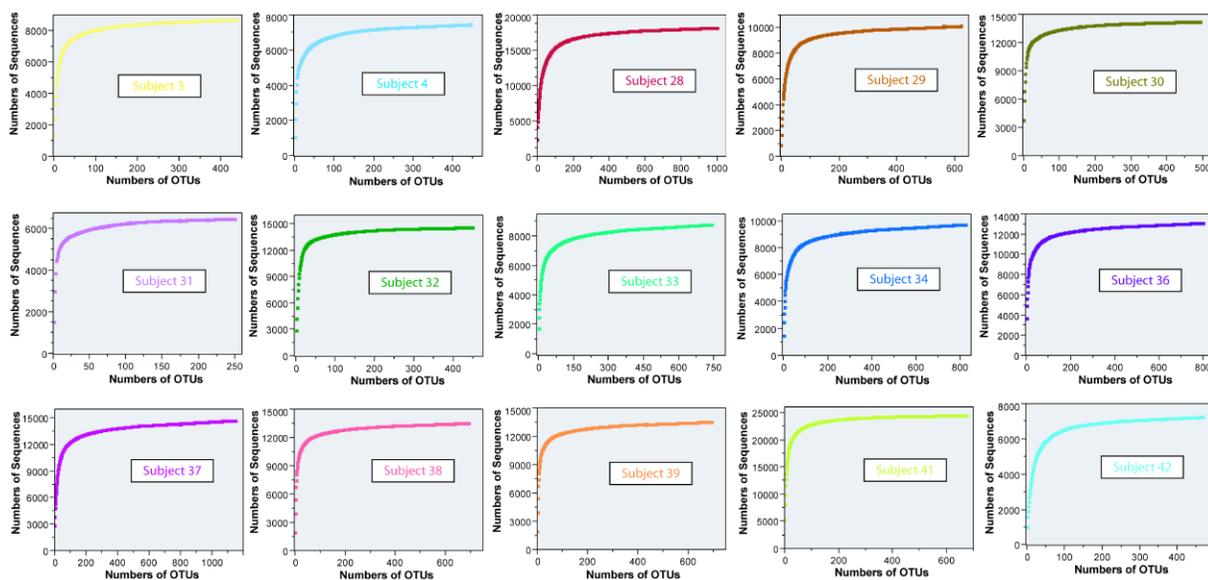
## Supplemental Figures



**Supplemental Figure 1. Example of Automated Ribosomal Intergenic Spacer Analysis.** An ARISA experiment showing changes in the microbial community during the initial baseline stage, when subjects have entered the hospital are placed on a controlled diet and a choline depletion stage when subjects are placed on a low-choline diet.



**Supplemental Figure 2. Rarefaction curve for all samples.** Rarefaction curve shows that sampling saturates the sequence space. The red arrows indicate the numbers of OTUs at which 93% (566 OTUs) and 95% (843 OTUs) of total sequences are under the curve.



**Supplemental Figure 3. Rarefaction curves by patient across samples.** Curves illustrate the differing levels of saturation.