MICROBIAL CONTRIBUTIONS TO DISEASE PHENOTYPES

by

Jonathan Ward McCafferty

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

2013

Approved by:

____
Dr. Anthony A. Fodor

____
Dr. ZhengChang Su

____
Dr. Shannon D. Schlueter

____
Dr. Todd R. Steck

©2013
Jonathan Ward McCafferty
ALL RIGHTS RESERVED

ABSTRACT

JONATHAN WARD MCCAFFERTY. Microbial contributions to disease phenotypes. (Under the direction of DR. ANTHONY A. FODOR)

The unseen world of microbes has a profound affect on everyday life. Complex microbial communities play a role in everything from climate regulation to human health and disease pathogenesis. Advancements in the field of Metagenomics are providing a window into the world of microbial communities with an unprecedented resolution. Next-generation sequencing technology is allowing researchers to describe the relationships between these complex microbial communities and their host environments. The research in this dissertation investigates these complex microbial host relationships and the various tools and techniques needed to conduct metagenomic research.

Chapter 1 presents a current overview of techniques at the disposal of researchers conducting metagenomics experiments. Topics discussed include qualitative DNA fingerprinting techniques, comparison between Next-generation sequencing platforms, and how to handle statistical analysis of large metagenomic datasets. Chapter 2 deals with the development of Peak Studio, a platform independent graphical user interface, intended to be a pre-processing tool for researchers conducting DNA fingerprinting experiments. Chapter 3 explores how time and microenvironment influence the structure of gut microbial communities in a mouse model. Two experimental cohorts of mice are analyzed through the use of Illumina HiSeq sequencing of the 16S rRNA targeted V6 hypervariable region. Also considered are the effects over time of inoculating mice with a founder microbial community. In total, this dissertation emphasizes the importance of experimental design and the development and use of technology in the exploration of complex microbial communities.

ACKNOWLEDGEMENTS

The research described in Chapter 3 was supported by NIH grant 2R01DK073338-05A1 (Dr. Christian Jobin). The Graduate Assistance Support Plan (GASP) through the Graduate School supported my graduate career at UNC Charlotte.

I would like to acknowledge the many contributions the members of the laboratory of Dr. Christian Jobin in the School of Medicine at the University of North Carolina at Chapel Hill made to Chapter 3. The Bioinformatics Service Division of the department of Bioinformatics and Genomics at the University of North Carolina at Charlotte, in particular Dr. Raad Gharaibeh and Dr. Wei Sha for their consultations and contributions to Chapter 3. Dr. Melanie Spencer, Dr. Robert Reid and Timothy Hamp, who were members of the laboratory of Dr. Anthony Fodor during my graduate career. They were invaluable members of the development team and assisted in the creation of the Peak Studio software (Chapter 2).

I would also like to acknowledge Dr. Christine Richardson, Department of Biology, University of North Carolina at Charlotte for her continuous support and guidance. My Doctoral committee members Dr. ZhengChang Su and Dr. Shannon Schlueter, Department of Bioinformatics and Genomics, Dr. Todd Steck, Department of Biology for the their advice and patients and to Dr. Anthony Fodor who took an interest in me and through his enthusiasm and thirst for knowledge guided, inspired and taught me to enjoy the journey that is science.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EQUATIONS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARISA	automated ribosomal intergenic spacer analysis
T-RFLP	terminal restriction fragment length polymorphisms
FSA	fragment analysis file
PCR	polymerase chain reaction
R	statistical software
SAS	statistical software
OS X	apple operating system
ABI	applied biosystems
JVM	java virtual machine
RAM	random access memory
GB	gigabyte
OTU	operational taxonomic unit
RDP	ribosomal database project
PCA	principal components analysis
PCoA	principal co-ordinate analysis
PC1	first principal component
WT	wild type
qPCR	quantitative polymerase chain reaction
DGGE	denaturing gradient gel electrophoresis

CHAPTER 1: HUMAN MICROBIOME ANALYSIS VIA THE 16S RRNA GENE

1.1 Abstract

The human associated microbiota has been linked to an ever-expanding set of diseases including obesity, cancer and inflammatory bowel disease. While the decreasing cost of sequencing is making whole-genome metagenomic shotgun sequencing more feasible, 16S rRNA based sequencing studies remain the most commonly utilized method to characterize a microbial community. In this review, we consider different methods to characterize a mixed microbial community by examination of the 16S rRNA gene. We discuss older, low-resolution methods such as Terminal Restriction Length Polymorphisms (T-RFLP) and Automated Ribosomal Intergenic Spacer Analysis (ARISA), which yield low-cost "snapshots" of the microbial community that can be generated rapidly. We next consider current high-throughput sequencing technology from 454 Life Sciences and Illumina. These techniques generate large amounts of data and careful consideration must be given to how low-quality sequences and PCR chimeras are removed from downstream consideration. We examine algorithms for clustering sequences into Operational Taxonomic Units (OTUs) and for assigning taxonomy. Finally, we consider methods for assigning statistical significance to differences between different microbial communities.

1.2 Introduction

Microorganisms exist in great abundance and inhabit virtually every conceivable environment on earth including the inside and outside of the human body. Environmental microbial communities range from the highly simplified community found within acid-mine drainage ecosystems [1, 2] to extraordinarily complex and diverse soil and ocean ecosystems [3-6]. Within the human microbiome, there is also a range of complexity ranging from the relative simplicity of vaginal samples to more complex habitats such as the human gut [7-11]. Prior to the application of sequencing technology to the study of microbial communities, our knowledge and understanding of microbial community composition had been limited to the subset of organisms that could be cultured. The sequencing technology that has driven recent advancements in culture independent molecular techniques is rapidly revolutionizing the field allowing for exquisitely detailed, yet low cost, descriptions of the structure of microbial communities. In this review, we discuss how technology has changed in the last few years, consider the informatics challenges that new sequencing technology is creating and examine recently developed solutions for these challenges.

1.3 Whole Genome Shotgun Sequencing Vs 16S Sequencing

All culture free identification methods begin with isolation of microbial DNA. The methods used to isolate DNA can have a profound impact on the observed microbial community structure [12, 13]. It is important, therefore, that in a series of experiments that the method of DNA isolation is not changed.

Once DNA has been isolated, two distinct techniques can be used to characterize the metagenome (Fig. 1). PCR can be used to target one gene that is used as a “barcode” for taxonomy. For bacteria, the chosen gene is usually the 16S rRNA gene, which is among the most conserved genes across evolutionary space in the microbial genome. As an alternative to surveying a single gene, whole genome shotgun (WGS) is accomplished through random shearing of genomic DNA into smaller fragments then ligating the necessary platform specific adapters to the fragments prior to the sequencing reaction. Whole genome sequencing bypasses the PCR amplification step removing a potential source of bias. Research conducted by the Human Microbiome Project (HMP) demonstrates that functional analysis of genes from whole genome sequencing produces results that have less variability that 16S community profiling [8, 14]. Whole-genome metagenome shotgun sequencing experiments, however, have higher requirements for the amount of starting material and are more sensitive to host contamination. Moreover, interpretation of shotgun sequencing experiments requires more sequences and hence whole-genome shotgun sequence datasets are more expensive to produce and require more time and computational resources to analyze. The Global Ocean Survey, at the time of publication in 2007 by far the largest whole genome metagenome shotgun experiment ever performed, consisted of 7.7 million Sanger sequencing reads for a total of 6.3 gibabases spread across 44 samples. Typically, a current sequencing strategy for whole-genome shotgun sequencing experiments will attempt to generate on the order of 2-10 gigabase of total sequence per sample from a paired-end generating technology. Because of this requirement for a large number of sequences, whole-genome shotgun sequence characterizations are usually done on the Illumina HiSeq platform, where the library preparations plus sequencing costs can potentially run into the hundreds of dollars per sample. By contrast, because 16S rRNA amplicon sequencing targets only a single gene, many fewer sequences are required to be informative. Before the advent of next-generation sequencing, a typical strategy would involve creating clone libraries from PCR products and picking and sequencing on the order of one hundred clones [15, 16]. With 454 sequencing, read-length decreased, but a typical study would generate on the order of thousands of sequences per sample [17, 18]. Recently, protocols have been published that describe the application of the Illumina HiSeq platform to 16S amplicon sequencing [19]. 16S rRNA sequencing based on Illumina HiSeq easily allows for tens of thousands to millions of sequences per sample [20, 21].

To economically run multiple sequences under either Illumina or 454 technologies, a “barcode” system is typically utilized in which a small DNA identifier (typically on the order of 4-10 nucleotides) is inserted adjacent to the 16S sequence in the PCR primer [22]. One lane of a typical paired-end HiSeq Illumina run may produce over 100 million sequences and cost on the order of $2,500US. By introducing bar-codes into both the 5’ and 3’ primers, the over 100 million sequences can be split between on the order of 100 to 1,000 samples allowing for a per-sample cost of as low as a few dollars. Per sample costs for 16S amplicon sequencing, therefore, remains much lower than for shotgun sequencing. Moreover, because fewer sequences are generated, the downstream analysis times and computational requirements are also significantly smaller.

Because of the high degree of individual variation in the human microbiome [7-11], in clinical studies large sample sizes are often required to ensure adequate power. Longitudinal study designs that sample patients across multiple time points can capture variation across time, allowing each patient to in some sense to serve as their own control. This approach can be crucial for detecting important changes against the backdrop of individual variation [18, 23]. Sampling many patients frequently across time can easily generate large numbers of samples, and the lower cost and simpler analysis path for 16S sequencing therefore offers a crucial advantage over whole-genome shotgun sequencing. However, straight-forward methods that allow for functional gene analysis [24] and taxonomy [25] of shotgun sequence datasets have recently been developed and as sequencing continues to approach being "free", the effects of the cost differences between 16S and shotgun methods will become less significant. In the future, therefore, we may see whole genome shotgun sequencing become the dominant method. For the immediate present, however, cost and data management concerns usually make 16S sequencing the default method for large clinical studies. A feasible strategy is to 16S sequence a large set of samples and then choose a subset of samples for in-depth whole-genome shotgun characterization.

1.4 Low Resolution Community Profiling

This barcode strategy described above for 454 and Illumina 16S sequencing has the disadvantage that samples need to be archived until a sufficient number of samples are available to make economical use of all of the sequences in the next-generation run. There are situations, however, when more immediate results are required. For example, in animal studies it may be required to know that the microbial community has achieved stability after a treatment before performing an intervention or terminating an experiment. In cases such as these, microbial "profiling" techniques, which were state of the art in the 1990s, may still have utility. Rather than directly sequencing the DNA sequence associated with the 16S gene, these profiling techniques use either a fluorescent tag incorporated into one of the PCR primers or a denaturing gel to separate DNA from different taxa (Fig. 2). Changes to the “profile” of the DNA can be used to detect changes to the overall microbial community, although it is usually not straightforward to identify which taxa make up the profile.

The three most popular DNA fingerprinting techniques are ARISA (Automated Ribosomal Intergenic Spacer Analysis) [26], T-RFLP (Terminal Restriction Length Polymorphism) [27], and DGGE (Denaturing Gradient Gel Electrophoresis) [28] (Fig. 2). T-RFLP performs a PCR targeting the 16S rRNA gene followed by application of one or more restriction enzymes. In ARISA, the PCR is performed in the region between the 16S and 23S gene. Since the number of nucleotides in this region is different for different taxa, changes in the microbial community will produce a distinct ARISA profile. In both ARISA and T-RFLP, a fluorescent primer is incorporated into one of the PCR primers and, typically, a Sanger sequencing machine is used to separate the DNA regions of different length (Fig. 2). As an alternative that does not require access to a Sanger sequencing machine, DGGE performs a PCR on the 16S gene and then uses a denaturing gel to separate DNA based on melting temperature. Different bacteria have different GC content and thus changes to the microbial community can be detected as changes to the DGGE profile.

In general, if access to a Sanger sequencing machine is available, T-RFLP and ARISA are easier to perform than DGGE and can generate results quickly. It is possible to perform DNA isolation and then generate interpretable T-RFLP and ARISA results within a 24-hour period for a cost on the order of one dollar per sample. Software, such as Peak Studio [29] can be used to interpret the results of T-RFLP and ARISA experiments.

While they are not a substitute for sequencing experiments, fingerprinting techniques are still an inexpensive and quick way to identify differences in microbial communities, and are currently in use by microbiology laboratories around the world, [4, 30, 31]. These techniques however are limited in their ability to provide taxonomic identifications. Fingerprinting techniques are especially useful as a way to test the success of DNA microbial isolation methods on difficult samples and can be used to troubleshoot samples of interest prior to sequencing.

1.5 Next Generation Sequencing

Before the advent of next-generation sequencing, Sanger sequencing [32] had been the dominant method for investigating microbial communities. While generating longer read lengths than currently popular next-generation platforms, Sanger sequencing suffers from several limitations including the requirement of building bacterial clone libraries, which had the potential to bias against genes that were harmful to the *E. Coli* that hosted the library. Moreover, the cost of Sanger sequencing inevitably limited the depth of sequencing making it more difficult to categorize the low-abundance members of the community. By eliminating the laborious clone library step, next generation sequencing experiments, in addition to being much cheaper, are much easier to perform and have the potential to be less biased than Sanger sequencing. As a result, few papers are currently published that make extensive use of Sanger 16S sequencing. The two next generation sequencing (NGS) platforms developed by Roche/454 Life Sciences (released in 2005) and Illumina/Solexa (released in 2007) [33], have ushered in a revolution in microbial ecology studies. While 454 pyrosequencing and Illumina currently produce shorter read lengths compared to Sanger sequencing, making alignment and de novo assembly of whole-genome shotgun reads more difficult, both platforms continue to improve. Initially read lengths from a 454 run averaged around 100 bases, increasing to 400 – 700 bases in just a few years while reducing overall cost to about 10 dollars per sequenced mega-base [34]. Illumina has also demonstrated the ability to quickly make adjustments to sequencing technology by increasing read length from 36 bases to currently greater than 100 bases and bringing down the cost to 12 cents per sequenced mega-base [34]. Read lengths will continue to grow with biochemistry refinements and advancements with microfluidics that will increase the speed of the sequencing reaction [35].

In performing a 16S rRNA sequencing experiment, an early choice that must be made is which region of the 16S rRNA gene should be sequenced. While the 16S rRNA gene is among the most conserved genes in bacterial genomes, it contains 9 hypervariable regions (V1-V9) that show considerable diversity among bacteria but can be targeted and amplified to facilitate identification down to the genus and species level [35, 36]. The relatively short read lengths of 454 and Illumina do not allow for sequencing across the entire 16S rRNA gene requiring an explicit choice of which variable regions are targeted when these sequencing platforms are used. In an early survey that used the GS20 platform on 454 sequencing, the ~100 base pair read length of that platform dictated the choice of the V6 primer [37]. As 454 sequencing matured and the read length approach 400 basepairs, more studies targeted regions within V1-V5, as bioinformatics studies suggested that these regions had improved taxonomic resolution [38, 39]. With the application of the Illumina HiSeq to 16S datasets, the V6 region has remained attractive [19, 20] as a 100 base-pair paired end read can sequence the entire V6 amplicon at 2X coverage. With 100 base-pair Illumina reads, the V4 region has also been targeted [21] likely allowing for greater taxonomic resolution [39] but at a cost of only partial overlap if a paired end approach is taken. As Illumina read-lengths increase, we can anticipate that future studies will more frequently target the V1-V3 and V3-V5 regions with a paired-end approach that will reduce the effect of sequencing error on downstream analyses.

1.6 Potential Sources Of Error And Preprocessing

Environmental deep sequencing of PCR amplicons using NGS technology enhances the ability to detect the low abundant members present in the community, what has been termed the “rare biosphere” [37, 38, 40, 41]. The same NGS technology that grants this unprecedented look at complex microbial communities also contributes to a possible overestimation of diversity due to the generation of low frequency error prone reads [40, 41]. Error rates using 16S rRNA amplicons can have a great effect on diversity estimates because every read in a 16S rRNA sequencing run is treated as a unique identifier for a member of the microbial community leading to inflation of diversity estimates [41]. PCR amplification can be a prime source of error bias and chimera formation in 16S datasets. The limitations of the next generation sequencing platforms are also sources of potential error accumulation and bias. Chemistry used in the 454 sequencing platform lacks a terminating functional group, allowing for the incorporation of multiple bases during a single injection cycle making an accurate assessment of the number of nucleotides in homopolymer region difficult. Huse and colleagues estimated that the errors involving homopolymer regions accounted for 39% of all errors using the GS20 454 sequencing platform on the V6 hypervariable region of

16S rRNA with insertions being the most common followed by deletions, ambiguous bases and substitutions [42]. While the Illumina platform uses bridge amplification instead of the emulsion PCR used by the 454 platform it also has limitations that produce potential errors. Minoche and colleagues report that Illumina sequences exhibit preferences for certain substitutions with a measurable GC bias demonstrated [43]. Any errors that occur during the sequencing process can have dramatic and profound effects on downstream analysis and therefore must be accounted for to prevent any false positive calls during taxonomic assignment procedures.

Errors accumulated during the sequencing process, regardless of platform selected, lead to artificially inflated estimates of diversity affecting the composition of the “rare biosphere”. Preprocessing to correct for these errors usually involves a quality filter step followed by a clustering algorithm to generate a set of Operational Taxonomic Units (OTUs) used for analysis.

The appropriate QA/QC pipeline to use on next generation sequencing data has been a source of considerable interest in the literature. Initially it was believed that filtering out reads with ambiguous bases, reads that contained an error in the primer sequence, or any reads that were too long or too short would be sufficient in minimizing error rates [42]. Kunin and colleagues argued, however, that even downstream of these QC filtering steps that diversity estimates from 16S rRNA 454 pyrosequencing studies can be inflated by two orders of magnitude due to sequencing errors [41]. By setting stringent thresholds for quality filtering and base trimming of reads and clustering at no greater than 97% they were able to eliminate most spurious OTUs in a library that sequenced a single *Escherichia coli* reference template [41]. In an alternative approach to eliminating spurious OTUs, an algorithm called PyroNoise [44] circumvented the 454 sequencing platform base calling algorithm and instead analyzed the underlying flowgrams produced by the sequencing machine. While potentially more accurate than competing methods, the pyronoise algorithm is computationally very expensive to run and can only be applied to 454 datasets.

To address the question of how data processing and error rates can affect diversity estimates in the rare biosphere; Huse and colleagues set out to analyze how different combinations of filter and clustering techniques influence the estimates of diversity [40]. By exploring pipelines that relaxed a requirement that no two sequences within an OTU had a greater divergence than the threshold of the OTU, Huse and colleagues demonstrated that spurious OTUs could be eliminated in a matter that was less computationally expensive than PyroNoise [40]. Using this method Huse et al argued that while previous analysis of the rare biosphere contained over estimates of diversity, the rare biosphere was not made up entirely of spurious OTUs [40]. The question of separating sequencing error from rare OTUs remains an area of active research.

One way to deal with sequencing error in the rare biosphere is to choose a clustering algorithm such as AbundantOTU that deliberately excludes the rare biosphere [45]. By using a recruitment strategy that builds consensus sequences from individual 16S reads, AbundantOTU takes advantage of redundant sequence information to achieve efficient run times. In our lab, we have found that AbundantOTU can cluster 100 million 100 base pair V6 Illumina 16S sequences in approximately 12 hours on a single CPU. Reads that are not recruited to consensus sequences by AbudantOTU represent sampling from rare species or error prone reads from abundant species. These left over reads can be used for further analysis but should be done with caution as they may be the source of diversity inflation [45]. In a recent study [7], it was found that the majority of reads from the Human Microbiome Project that failed to be clustered by AbundantOTU were in fact chimeric as detected by UCHIME [46]. This suggests that reads that are not incorporated by AbundantOTU are frequently the result of error.

1.7 Chimera Detection

16S datasets characterized by next-generation sequencing requires an initial amplification of sample through PCR. Chimeric sequences are artifacts in the PCR process that result in the formation of product that is the combination of two or more parent sequences. Anomalies is sequences from diverse origins have been identified in public repositories creating the appearance of novel non-existent organisms [47]. In 2005, it was estimated that the error rate of sequences in public databases is 5% with chimeras representing the majority of anomalies [48]. Chimera detection is an active area of research with many researchers developing algorithms to filter sequences and limit the introduction of chimeras into analysis pipelines.

Early approaches to chimera detection utilized a comparison of calculated evolutionary distance with that of the known rate of variability in the 16S rRNA gene with highly divergent sequences flagged as chimeras [48]. Chimera Slayer was developed to address the short read lengths and large data sets produced by NGS sequencing platforms [49]. Chimeria Slayer uses a multiple sequence alignment of a chimera free reference database that can be searched by query sequences to identify potential chimeras. Edgar and colleagues have recently developed what is likely to date the most sensitive and accurate chimera detection software program UCHIME [46], which can work either by mapping sequences to a reference database or in a “de novo” mode that does not require a reference database [46]. UCHIME has demonstrated an increase in speed and sensitivity compared to the next best chimera detection algorithm, Chimera Slayer, while preserving lower error rates [46].

1.8 Taxonomy Assignment

Accurate taxonomic assignment of high throughput sequencing data is essential to our understanding of the structure and composition of microbial communities and defining ecological roles played by community members. Without taxonomic information findings about communities cannot be related to known attributes of microbes at varying levels of resolution [50]. A principle challenge is obtaining accurate assignments using the shorter reads produced by next generation sequencing. A common method for evaluating the taxonomy of a sequence is to simply BLAST the sequence against some reference database. However, since many sequences in reference databases are annotated simply as “uncultured organism”, and the query sequence can match many reference taxa with sometimes conflicting taxonomic annotations, this method often leads to unsatisfying results.

Classification algorithms attempt to assign taxonomy to sequences in a more systematic matter than is achievable through a simple BLAST search. Arguably the most widely used taxonomic classifier The Ribosomal Database Project (RDP) algorithm classifies taxa based on the co-occurrence of 8-mers in a query sequence and a reference database. Trained on Bergey’s Taxonomic Outline accuracy of classification can be seen down to the genus level for near-full-length and 400 base pair partial rRNA sequences [51]. Shorter 200 base pair partial sequences were accurate down to the family level

[51]. Given the simplicity of the algorithm used by RDP classification scheme, its high level of accuracy is perhaps surprising. Misclassifications are primarily caused by errors present in the underlying training set of reference sequences, but in the case of shorter reads a lack of information contained in the sequence could lead to misclassifications [51]. Because it is based on 8-mers, which can be indexed for rapid retrieval independent of the size of the reference database, the RDP algorithm is extremely computationally efficient, a factor that undoubtedly contributes to its enduring popularity. Well regarded alternatives to the RDP algorithm include techniques based on the Greengenes [52] and Silva [53] databases. A recent paper demonstrated new methods that have led to an improved Greengenes taxonomy [54]. Taxonomic classification remains an active area of research as well as a source of much debate and controversy and we can expect continued refinement of taxonomies and classifiers as more datasets become available and algorithms continue to improve.

Whichever classifier is used, query read length is a contributing factor in correctly assigning taxonomy. A choice that must be made in an analysis pipeline is whether to directly classify the short reads produced by next generation sequencers or to map those reads to full length 16S rRNA sequences and instead classify the full-length references. Using the RDP classifier, increasing query reads from 50 bases to full-length 16S rRNA will generate a greater than 5% increase in accuracy at the Phylum level and a greater than 39% increase in accuracy at the Genus level [51]. With short reads (such as the ~100 basepairs produced by early 454 technology or recent Illumina HiSeq technology), classifying reference OTUs instead of the reads directly is therefore clearly attractive. Individual sequence reads (or consensus or representative sequences from OTUs) can be mapped to reference databases with simple best hits from blast searches or from methods such as GAST [55] or align.seqs from Mothur [56] that consider global alignments. A study [7] from the Human Microbiome Project demonstrated that nearly every taxa in the HMP collection was previously observed as a full length sequence in the Silva database. This makes the strategy of mapping short-read sequences to a full length database a feasible option for human metagenomic studies, with the obvious caveat that this approach will be unable to discriminate two taxa that play biologically distinct roles but have identical sequences within the sequenced region.

1.9 Statistical Analysis

In order to understand how the microbial community contributes to health and disease phenotypes, it is necessary to perform inference in order to assign probabilities with which to reject null hypotheses that the state of the microbial community is not associated with subject characteristics. A straightforward approach to this problem is to choose a taxonomic level (phyla, class, order, family, genus or OTU) and form a null hypothesis for each taxa that the taxa is not associated with the phenotype of interest. P-values for each null hypothesis can be generated by univariate statistical tests. For example, for a case-control experiment, the t-test can be used, or if the sample size is large enough, the Wilcoxon test in order to avoid the parametric assumptions of the t-test. This approach has been used numerous times in the literature [17, 18].

One possible limitation of this approach is that it will lead to over-fitting and spurious conclusions if a simple threshold of significance (for example $p < 0.05$) is used. This is because an independent test is run for each taxa in the experiment. If, for example, there were 1,000 OTUs in an experiment (a not atypical number for a human gut survey) and a simple p-value threshold of p <0.05 were used, we would expect 50 significant "hits" even if completely random data were fed into the t-tests. In order to avoid over-fitting of data, appropriate correction for multiple hypothesis testing is required. One simple approach, Bonferonni correction, adjusts the p-value directly by dividing the p-value threshold by the number of tests that are being run. So if 1,000 null hypotheses are tested, the p-value that would be used as a threshold of significance is $5 * 10^{-5}$ (that is, 0.05/1000). The probability that any of the hits detected at this threshold of significance are false positives is 0.05. Bonferonni correction sets a rigorous threshold for interpretation of genomics experiments, but is often considered to be too conservative for genomics experiments. A popular alternative is false discovery rate (FDR) based metrics. At a 5% FDR threshold, we would expect 5% of the hits to be false positives. This is a far less stringent threshold than a Bonferonni corrected p-value of 0.05, in which there would only be a 5% chance that any of the hits would be false positives. Popular methods of calculating false discovery rate include Storey's q-value method [57] and the Benjamani and Hochberg false discovery rate method [58]. The Benjamani and Hochberg method in particular is very easy to calculate and has a straightforward interpretation. Given a list of p-values that result from a series of independent statistical tests, the list is sorted with the smallest p-values at the top. For each p-value, a corrected metric is calculated which is N*p/k where N is the number of null hypotheses that are being tested, p is the p-value produced by the independent statistical test, and k is the rank (the smallest p-value ranked 1, the next p-value ranked 2 and so forth). To determine what hits are significant at a 5% false discovery rate, one simply starts at the top of the list and continues until the N*p/k value exceeds 0.05.

An alternative to performing multiple statistical tests on metagenomic data is to reduce the high dimensionality of metagenomic data sets by finding individual metrics that describe the state of the metagenomic community and performing inference on those metrics. Popular examples of such a metric include diversity metrics that attempt to describe the complexity of the microbial community. The simplest measure of microbial community complexity is richness, which is simply the number of taxa present in a sample. In NGS experiments in which barcodes are utilized, there are inevitably very different numbers of sequences per sample and this can potentially skew richness methods. A simple but effective technique to correct for this is to randomly re-sample each sample a set number of times and report as richness the average number of taxa observed across the re-samples.

Another often-used metric is the Shannon diversity index. This measures diversity through a log proportionality of species abundance in each sample. To calculate Shannon diversity, each taxa in the sample is converted to a proportion (for example if 12% of the sequences were assigned to Firmicutes, p for Firmicutes would be 0.12) and the Shannon diversity is simply calculated as $-\Sigma p * \log p$ summed across all of the taxa. Shannon diversity is easy to compute but it has been argued [59] that it lacks a straightforward biological interpretation. Shannon diversity reflects a mixture of richness (as defined above) and evenness (how equally reads are distributed across the taxa). A high Shannon diversity, therefore, can reflect either high richness or high evenness. Directly reporting richness and evenness rather than Shannon diversity may lead to results with a more straightforward biological interpretation.

An alternative to diversity metrics is to find single variables that describe the entire microbial community. Microbial ecologists have long utilized multivariate statistical analysis as a way of visualizing and explaining diversity patterns based on environment, time, geographical location, or disease states in high dimensional data sets. Principal component analysis (PCA) and principal coordinates analysis (PCoA) are two often used metrics for identifying patterns in metagenomic data. Both techniques are in a class of unsupervised statistical models that compress high dimensional data into a set of new variables that will explain the variance contained in the data in a lower dimensional space. While PCA and PCoA share similar assumptions and objectives in that they project the similarities between samples onto a new coordinate system, the input matrix used and data interpretation differs [60]. Standard implementations of PCA are conducted with covariance or correlation matrices [60]. In contrast to PCA, PCoA can use any distance matrix as input. In microbial ecology UniFrac is a popular distance metric used to analyze microbial community data sets [61]. A UniFrac distance is calculated between any two samples by constructing phylogenic trees. Environment similarities are determined through the distance metric based on the number of shared branch lengths in the phylogenetic tree. Weighted UniFrac, a modification to UniFrac incorporates abundance information into the branch length calculation in order to track changes in community organism populations [61]. While UniFrac is currently a popular choice for microbial community studies, other distance metrics have been implemented with comparable results. A recent study demonstrated that 18 distance metrics obtained broadly similar results in a study of an elderly Irish cohort [62]. A metric long popular with ecologists due to its simple calculation and ease of interpretation is the Bray-Curtis dissimilarity. This metric is not a true distance metric but works by quantifying dissimilarity between two samples based on the count of common taxa divided by the total number of taxa present.

Canonical unsupervised statistical tests ask whether changes to the microbial community are statistically associated with changes to phenotypes of interest. An alternative analysis path, supervised classification, instead asks whether the state of the microbial community can predict phenotypes of interest [63]. One goal of this type of analysis is to identify groups of microbes that can be used as markers for disease and distress. In supervised classification, models are constructed from a set of training data with categorical information, case and control for example, and then when new unlabeled data is introduced the model makes a prediction as to which category the new data belongs. The field of machine learning offers many models that could potentially reveal relationships between metagenomic data and host phenotypes [63]. Random forests (RF) classifiers work by generating decision trees from a random subset of available features and then discriminating between categories by choosing the maximum number of category predictions. RF's have been applied to characterize metagenomic signatures [64] but depending on the data are not always the best classifier choice [65]. Other models that have successfully been applied to metagenomic data included Elastic Net (ENET) [66] and a technique that combines k-nearest neighbor and Support Vector Machines (SVM) [67].

Overfitting is always a concern when dealing with predictive models. A supervised classification model is trained on a set of known data and the more complex the data is the more the model is prone to describing the noise in the data rather than the underlying relationship. This causes the model to be highly accurate on the training set but to falter when new data is presented. Cross-training validation where models are repeatedly trained on a subset of data and then tested on the “left out” portion of the data are routinely used as tests for overfitting. Waldron and colleagues [66] demonstrated, however, that even this sort of approach is not a guarantee of preventing model over-fitting and they highlight the importance of using datasets that were in no way used in model building steps in order to test the model to avoid generating irreproducible results.

1.10 Conclusion

The advent of next-generation sequencing is spawning a revolution in microbiology allowing for the analysis of whole communities rather than only organisms that can be cultured. While the potential applications of this technology seem limitless, as sequencing becomes less expensive the costs and efforts associated with data analysis become an ever-larger part of the budget of sequencing experiments [68]. Even though 16S datasets are substantially simpler and smaller than whole-genome metagenomic datasets, careful attention must be paid to pre-processing, clustering, taxonomy and statistical techniques if reproducible results are to be obtained from 16S datasets. Fortunately, popular software suites including Qiime [69] and Mothur [70] collect pre-processing, clustering and analysis packages allowing for application on these methods by users with minimal requirements for scripting or coding by the end user. While there is no single “correct” pipeline for analysis of 16S data, a strong grasp on fundamental statistics and a good understanding of how the algorithms in the chosen pipeline work are essential to avoiding costly errors that will lead to irreproducible results. Biologists who lack a background in these areas should strongly consider collaborations with experts in bioinformatics and statistics. As the Human Microbiome Project [8] has demonstrated, such multi-disciplinary teams can make substantial and exciting progress in linking the structure and function of metagenomic communities to human health and disease outcomes.

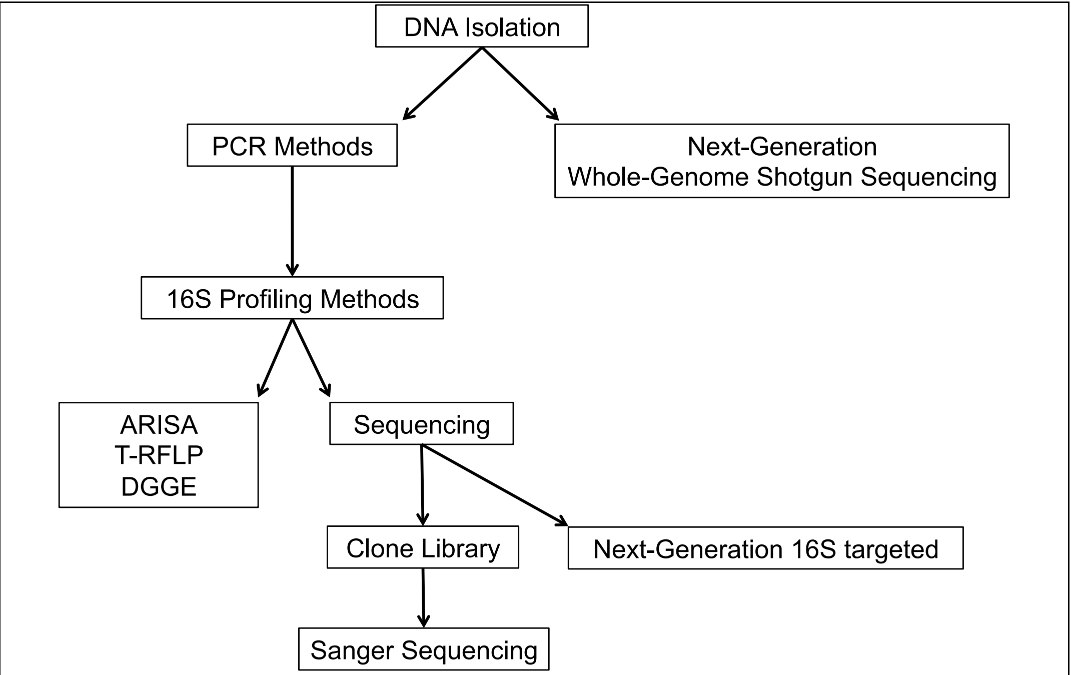


FIGURE 1.1: Flow chart illustrating possible techniques for microbial community analysis.

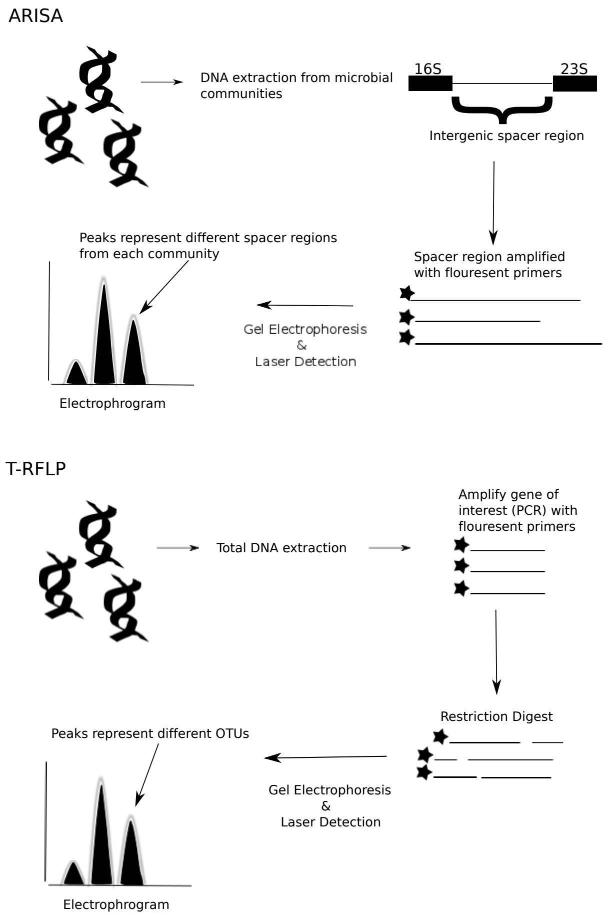

ARISA
DNA extraction from microbial communities
16S
23S
Intergenic spacer region
Spacer region amplified with flouresent primers
Gel Electrophoresis & Laser Detection
Peaks represent different spacer regions from each community
Electrophrogram
T-RFLP
Total DNA extraction
Amplify gene of interest (PCR) with flouresent primers
Restriction Digest
Gel Electrophoresis & Laser Detection
Peaks represent different OTUs
Electrophrogram

FIGURE 1.2: Typical experimental flow for ARISA (top panel) and T-RFLP (bottom panel). DNA is extracted from microbial community samples. In the case of ARISA the intergenic distance between the 16S and 23S rRNA gene is measured while T-RFLP uses fragments from a restriction digest of a gene of interest (typically the 16S rRNA gene). Both techniques produce an electrophrogram where different members of the sample community are represented by peaks.

CHAPTER 2: PEAK STUDIO: A TOOL FOR THE VISUALIZATION AND ANALYSIS OF FRAGMENT ANALYSIS FILES [29]

2.1 Abstract

While emerging technologies such as next generation sequencing are increasingly important tools for the analysis of metagenomic communities, molecular fingerprinting techniques such as Automated Ribosomal Intergenic Spacer Analysis (ARISA) and Terminal Restriction Fragment Length Polymorphisms (T-RFLP) remain in use due to their rapid speed and low cost. Peak Studio is a java based graphical user interface (GUI) designed for the visualization and analysis of fragment analysis (FSA) files generated by the Applied Biosystems capillary electrophoresis instrument. Specifically designed for ARISA and T-RFLP experiments, Peak Studio provides the user the ability to freely adjust the parameters of a peak-calling algorithm and immediately see the implications for downstream clustering by Principal Component Analysis. Peak Studio is fully open-source and, unlike proprietary solutions, can be deployed on any computer running Windows, OS X or Linux. Peak Studio allows data to be saved in multiple formats and can serve as a pre-processing suite that prepares data for statistical analysis in programs such as SAS or R. Source code binaries, a user manual and demonstration videos are available at www.fodorlab.uncc.edu/PeakStudioPage.html.

2.2 Introduction

Describing the diversity and complexity in mixed microbial communities is essential to understanding the role microbes play in an environment. Microbiology techniques utilizing PCR are well established [71] and provide a way to directly amplify microbial genes from samples, removing the need to culture [72]. Molecular fingerprinting techniques such as the Automated Approach for Ribosomal Intergenic Spacer Analysis (ARISA) [26] and Terminal Restriction Fragment Length Polymorphisms (T-RFLP) [27], provide a cost efficient and time effective way to produce microbial community diversity profiles. Numerous studies have demonstrated the reproducibility and accuracy of these techniques in microbial communities ranging from terrestrial [73], to aquatic [74] to the human microbiome [18]. While advancing technology is continuing to increase the affordability and popularity of next generation sequencing techniques, ARISA and T-RFLP remain proven microbiology techniques that still maintain an advantage in cost and speed over next generation sequencing and should not be overlooked in the scientific toolbox. Because T-RFLP and ARISA profiles can be generated quickly and for low cost, they remain useful as a quality-control step to check DNA extraction prior to techniques such as 454 sequencing, where per sample costs can be as much as one to two orders of magnitude higher.

Because of their low cost and speed, it is relatively easy to generate datasets with hundreds of T-RFLP and ARISA samples with modest experimental efforts. The visualization and management of such datasets is challenging, in part because the software that ships with the ABI sequencing machine is not open source and may not be freely distributed to client computers. Moreover, peak calling algorithms are notoriously unreliable and biologists often wish to have direct control over how peaks are called in their spectra. Peak Studio was designed to provide a data-browsing interface that allows the user to see in real time the consequences of changing which peaks are called for downstream clustering methods. Other software tools allow users to analyze peak patterns including a web-based interface called T-REX [75], T-RFLP Stats which is a downloadable package of scripts run from a command line environment [76], Ribosort an R package for analysis of microbial community profiles [77] and a clustering algorithm written in R designed for DNA fingerprinting studies [78]. These valuable tools utilize as input a data table exported by ABI's GeneMapper® software as input. Peak Studio allows the user to generate the input required for these and other statistical tools (such as R and SAS) by directly accessing the binary fragment analysis files (FSA files) generated by the ABI instrument. By allowing for full control over how peaks are called from within an interactive browsing environment, and permitting the export of the resultant peak tables in a variety of text file formats, Peak Studio can be easily incorporated into existing data analysis pipelines while giving users more control over data pre-processing.

2.3 Input And Visualization

Peak Studio is designed to accept FSA files generated by the ABI capillary electrophoresis instrument. The user has the option to select the appropriate colors corresponding to the dye colors used when samples were run on the ABI machine. The spectra will display in the primary window with the peaks identified by their respective dye color. Once the files are loaded, a fully sortable and modifiable table displays information on each spectra, including name, dye color, and current peak color. The user has the freedom to modify the color of called peaks and non-peak regions to best suit their viewing or organizing needs. Any metadata (i.e. user annotations on samples) that is added will also appear in the table as additional columns. Peak Studio provides continuous scrolling and zooming in real time.

2.4 Data Analysis

Peak Studio allows for PCA clustering within the program or for export of binned data in tabular format suitable for analysis by full statistics packages such as R. A sizing table is also available for export that can be utilized by other available statistical software such as T-REX [75], T-RFLP Stats [76] or Ribosort [77]. Ultimately, the number of FSA files that can be analyzed at one time is limited by the amount of RAM available to the Java Virtual Machine (JVM). We tested the software on both Mac OS X and Windows XP platforms and successfully analyzed several hundred FSA files allocating 1 GB of RAM to the JVM.

2.5 Peak Calling Heuristics

Accurate identification of peaks is a critical step in ensuring that the data is prepared for further analysis. Our peak-calling algorithm applies linear interpolation to separate signals of peaks from that of baseline. The algorithm works by using a configurable parameter set that contains thresholds for values such as slope and peak heights, assigning each data point to one of five possible phases: non-peak, peak, up-slope, down-slope or inter-peak. A peak is recognized as a collection of points that meet the requirement of beginning at an up slope phase and ending at a down slope phase. Taking the difference between the highest and lowest data points in the region containing the peak determines the height. If the newly detected peak fails to meet the user determined height threshold, the data that is contained in the phase is then relabeled as a nonpeak region. Adjusting the parameter set allows the user to redefine what constitutes a peak with the resulting peak calls seen in real time. Since any peak-calling algorithm has the potential of missing peaks or miscalling peaks, Peak Studio combines automated peak detection with the ability for the user to visually inspect and manually select peaks that need to be adjusted.

Analysis of a T-RFLP or ARISA file begins with assigning peaks to a standard ladder that is run with each sample. In order to determine whether peaks have been assigned correctly to the standards ladder, Peak Studio defines a QC score by taking a set of three peaks at a time and applying linear interpolation using the location of the outside peaks to predict the location of the center peak. The QC score is the sum of the absolute value of the difference between the predicated peak location and observed location divided by the number of total peaks called in the spectra. In our experience, this value is usually less than 1 basepair if the standards have been correctly assigned. Peak Studio has two modes that allow for assignment of standard peaks. In manual mode, peak assignment is carried out with adjustable parameters as above and the user can manually adjust the peaks if the QC score is high or if the resulting number of peaks is not in accordance with the expected number of peaks. In automated mode, peaks are assigned a probability of being correctly called (based on their magnitude) and peaks are removed or added until the correct number of peaks is achieved with a low QC scores (see the Peak Studio manual for more details). For most spectra, the automated mode will generate the correct assignment to the reference spectra, but through use of the manual mode of reference peak assignment, samples that have misidentified peaks can be salvaged allowing researchers access to data that may have been discarded by ABI's GeneMapper® software.

Correctly detecting peaks is a non-trivial problem often hindered by interference from high signal-to-noise ratios (SNR) in the spectra. ABI's GeneMapper® software incorporates baseline correction and data smoothing algorithms into the peak detection process. Peak Studio also implements optional baseline correction and smoothing and provides the user the ability to choose raw or smoothed baseline corrected data for downstream analysis. In Peak Studio, baseline correction is a two-fold process where an estimated baseline is generated then subtracted from the spectra to establish a baseline independently for each spectrum. Regression points for a continuous baseline are derived by employing a sliding window and recursive histogram approach [79], where the mean of the noise distribution in the window is recovered by the mode of the histogram. To further reduce spectral noise we use a Savitzky-Golay low-pass smoothing filter [80]. A minimum peak height threshold is also available for user adjustment in the parameter set, similar to the way ABI's GeneMapper® provides access through the analysis methods panel. The recommendation from Abdo et al is (2006) to set a height threshold as low as possible to maximize background noise; Peak Studio by default sets a height threshold of 25 fluorescent units for data spectra but this can easily be adjusted downward as appropriate.

2.6 Results

For demonstration purposes, we examined a data set involving the gut microbiome from fecal samples from human subjects on strictly controlled experimental diets over a two-month period in which choline was manipulated [18]. The study was split into three phases with a specific diet administered during the baseline phase, choline depletion phase and the choline repletion phase [18]. Longitudinal sampling of 15 female patients generated 74 samples [18]. ARISA was conducted on each sample and fragment analysis files generated by the Applied Biosystems capillary electrophoresis instrument were used as input into Peak Studio.

Evaluation of Peak Studio's peak calling heuristics was conducted by comparing ARISA data from sample 4B2_A07, which is from patient "4" in our study during the baseline phase before choline manipulation [18]. We compared this ARISA sample analyzed with the default AFLP (Amplified Fragment Length Polymorphisms) settings in ABI's GeneMapper® against the default settings of Peak Studio. Standard spectra (resulting from size standards) generally have a very high signal-to-noise ratio making peaks relatively easy to separate from background noise. Comparison of the area under the peaks between Peak Studio and ABI show a very high correlation (FIGURE 2.1A). Data spectra (resulting from 16S-23S intergenic lengths in our metagenomic samples) are inherently noisier and have a much lower signal-to-noise ratio increasing the difficulty in correctly separating peaks from background noise. Comparing the area under the peaks from the data spectra between ABI's GeneMapper® and Peak Studio, we still observe a reasonable correlation (FIGURE 2.1B), although not as high as for the standards spectra. We conclude that, at least for this human gut sample, the default settings in Peak Studio and GeneMapper yield broadly similar results.

Peak Studio gives the user the ability to select multiple samples to examine regions of interest that may be different between samples and display information about each peak (FIGURE 2.2). Principal Component analysis (PCA) can be initiated by user selection of the starting and stop locations, in base pair space, along the x-axis with a minimum height threshold for peaks to be considered. A new viewing window will display the PCA, colored by the users choice of Peak Color in the primary display, which can be adjusted in real time (FIGURE 2.3). The data matrix used to produce the PCA can be exported as a tab delimitated text file, also available for export is a sizing table in the same format used by ABI's GeneMapper® software. These files are compatible with other statistical analysis programs, such as SAS, R, T-REX [75], T-RFLP Stats [76] and Ribosort [77].

2.7 Discussion

Molecular fingerprinting techniques, such as ARISA, can be used to rapidly generate snapshots of microbial diversity. Often DNA extraction and generation of the ARISA or T-RLFP profile can be completed in a single day. Peak Studio was designed to be a user-friendly data browsing and visualization application that gives the user a fine level of control over peak calling and acts as a pre-processing step that works in concert with currently available statistical analysis software. While the emergence of next generation sequencing techniques is ushering in a new paradigm in microbial diversity studies, Peak Studio provides support for proven microbiology techniques that are still widely used.

2.8 Acknowledgements

We would like to thank Dr Katherine Lemon in the Division of Infectious Diseases, Children's Hospital Boston and the Department of Molecular Genetics, Forsyth Institute for helping us test the software, Dr Robert Kosara in the Department of Computer Science, College of Computing and Informatics, University of North Carolina

Charlotte for his insightful suggestions with data visualization, and Dr Michael Thomas Flanagan for making his code publicly available (www.ee.ucl.ac.uk/~mflanaga)**,** which we used for smoothing and area calculations in Peak Studio.

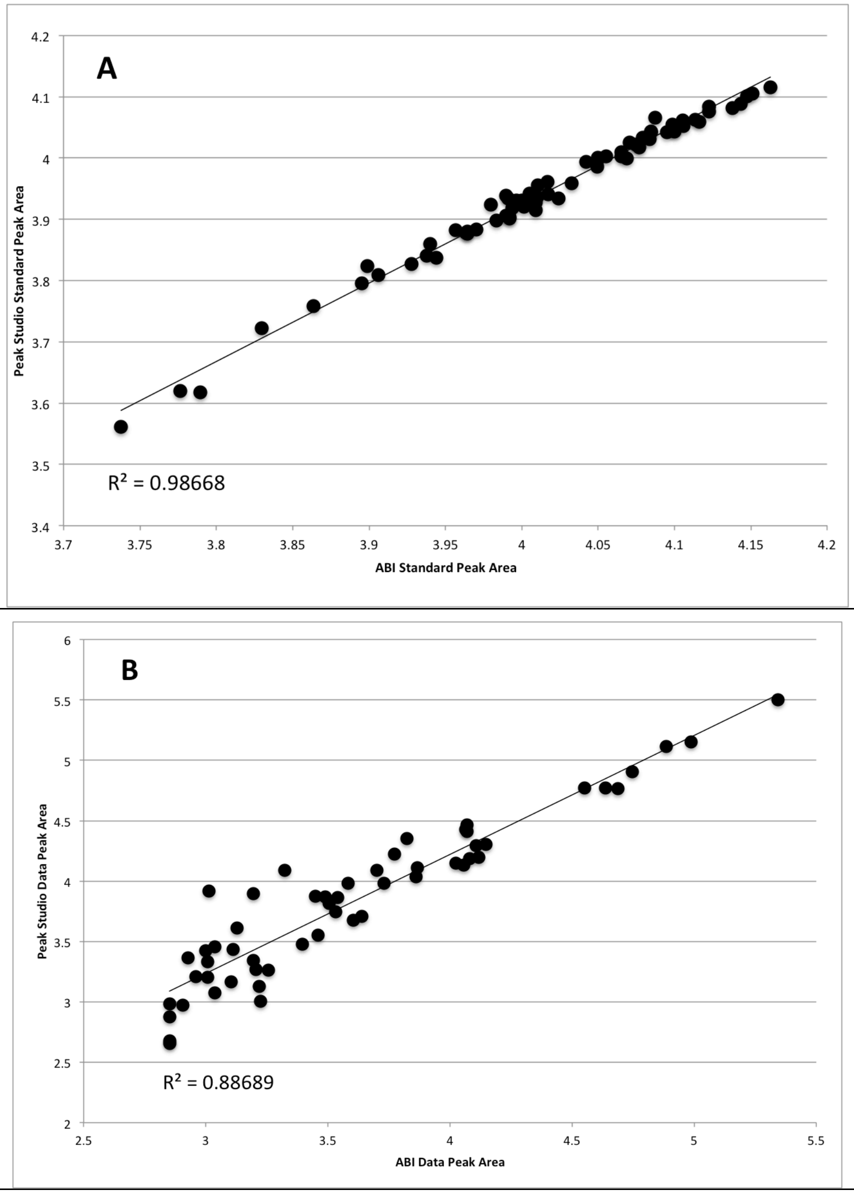
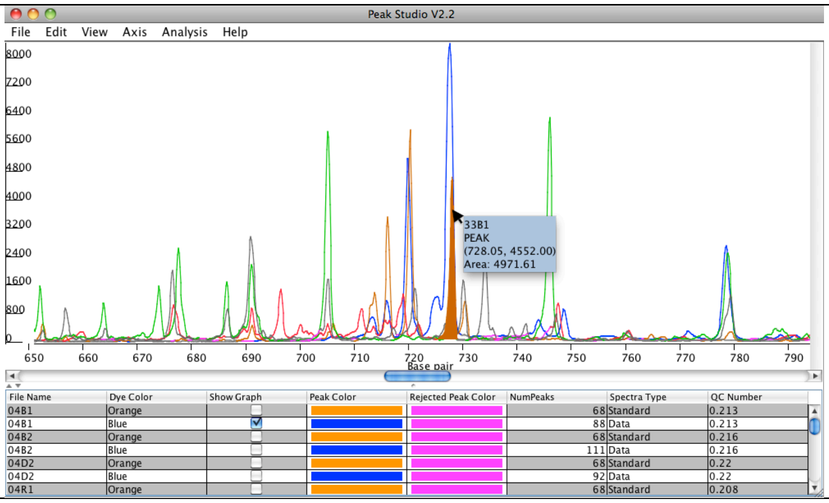


FIGURE 2.1: Linear regression of peak area of ARISA data from a human gut metagenomic sample 4B2_A07 [18] analyzed with the default settings in both Peak Studio and GeneMapper® for (A). Standards spectra and (B). Data spectra.



File Name	Dye Color	Show Graph	Peak Color	Rejected Peak Color	NumPeaks	Spectra Type	QC Number
04B1	Orange	☐	🟧	🟪	68	Standard	0.213
04B1	Blue	☑	🟦	🟪	88	Data	0.213
04B2	Orange	☐	🟧	🟪	68	Standard	0.216
04B2	Blue	☐	🟦	🟪	111	Data	0.216
04D2	Orange	☐	🟧	🟪	68	Standard	0.22
04D2	Blue	☐	🟦	🟪	92	Data	0.22
04R1	Orange	☐	🟧	🟪	68	Standard	0.208

FIGURE 2.2: Data spectra of multiple samples viewed at the same time. Resting the mouse on an individual peak will highlight the area under the peak and display an information window containing the sample name, what the algorithm identified it as, the xy coordinates and the area under the curve.

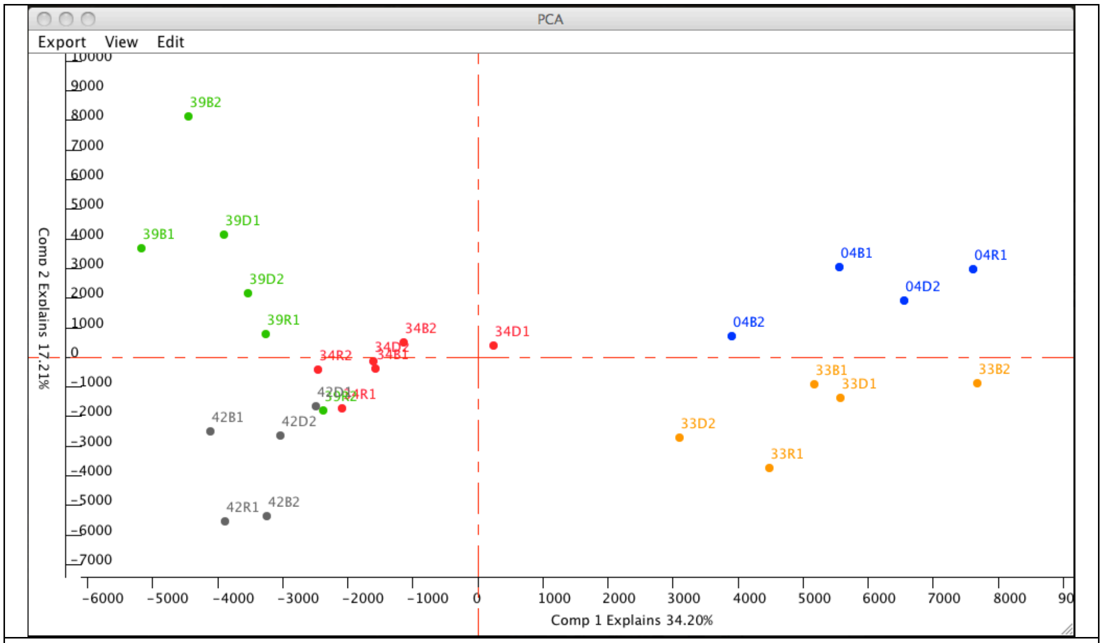


FIGURE 2.3: This dataset is a subset of samples from a study completed by Spencer et al in 2011. Patient sampling was conducted at various time-points indicating different diets throughout the study. Patients were subjected to strictly controlled diets used to monitor any fluctuations in microbial composition. Diets fell into different phases, baseline phase (B1, B2), choline depletion phase (D1, D2) and a choline repletion phase (R1, R2) [18]. Four character sample names identify both the patient and the current diet the patient was on at that time. For example sample 33B1 identifies patient 33 on diet B1. Viewing the result of the PCA reinforces what we would expect to see; samples taken from the same individual will cluster together because an established microbiome does not undergo a comprehensive change in response to short-term dietary modifications [18].

CHAPTER 3: STOCHASTIC CHANGES OVER TIME AND NOT FOUNDER EFFECTS EXPLAIN CAGE EFFECTS IN MOUSE MODELS OF THE GUT MICROBIAL COMMUNITY

3.1 Abstract

Cage effects in studies of the mouse microbiome are pronounced and can confound experimental design. We show that cage effects in animals removed from germ free conditions take several weeks to develop and are not mitigated by an initial gavage. This suggests that stochastic differences that develop over time in different cages, rather than initial founder differences between cages, are the cause of cage effects. Mice that are allowed to naturally acquire a microbial community from their cage, but not mice treated with gavage, show a cage effect in inflammation induced by DSS. This initial gavage influenced, but did not eliminate, a successional pattern we repeatedly observed in both Wild Type (WT) and Interleukin-10-deficient (*Il10*$^{-/-}$) mice in which Proteobacteria became reduced over time. Our results argue that the long term effects of gavage are subject to mitigation by cage and time, which must both be explicitly considered in the interpretation of microbiome mouse experiments.

3.2 Introduction

Since Darwin's formation of the theory of evolution based on observations made in the Galapagos Islands, island ecology has played a key role in our understanding of how communities form and respond to change. The mammalian gut can be thought of as an island inhabited by a complex assemblage of microbes. It has been demonstrated in humans that the initial micobiome in different body sites is undifferentiated and is set by mode of delivery [11]. Over time, selection pressure on the human microbiome sculpts microbes in each body site, so that, for example, the adult oral microbiota are largely distinct from the adult gut microbiota [81].

In the experiments in this study, we look at successional patterns over time in WT and *Il10*$^{-/-}$mice. Interleukin-10 (*Il10*) is an anti-inflammatory cytokine. Microbial commensals that are well tolerated in WT mice cause severe inflammation in *Il10*$^{-/-}$ mice [20, 82] suggesting the *Il10* gene plays a key role in host-micobial interactions. In a previous paper, we examined WT and *Il10*$^{-/-}$mice 20 weeks after they were removed from germ-free conditions [20]. We found that the proteobacteria *E. coli* was greatly expanded in the week 20 *Il10*$^{-/-}$ mice and that the *pks* island within the *E. coli* genome was essential in driving an inflammation phenotype to tumor formation in the presence of the carcinogen AOM. This observation led us to propose a "two-hit" model in which early inflammation allows expansion of bacteria with genotoxic potential and this expansion in turn disrupts host phenotype.

In this previous study, we noticed a strong cage effect in which animals within the same cage had similar microbial communities. To account for these cage effects, in the previous study we reported median values of each cage. In the present study, we wished to explore the causes and consequences of these cage effects and more fully consider a more formal statistical model to handle these cage effects. We therefore used Illumina sequencing to characterize the 16S gene from fecal samples collected over time for four cohorts of mice following removal from germ-free conditions. In one experiment, we added 2 week and 12 week time-points to our previously published (20 week) WT and

Il10$^{-/-}$ cohorts. In this experiment with WT and *Il10*$^{-/-}$ mice, we let the mice acquire their microbial community from the cage microenvironment. We report that time is the dominant force in structuring the microbial community but strong effects of genotype and cage are observable in our dataset.

Because by chance different cages might have different initial communities of microbes, and there was no other source of microbes for mice removed from germ-free conditions, we suspected that these different "founder" effects might explain the cage effects. Under this model, mice removed from germ free conditions would "amplify" whatever microbes they by chance initially encountered in their cages. To test this hypothesis, we performed a longitudinal time-series on WT mice in which the initial microbial community in one group of mice was set by gavage and another was allowed to acquire the microbial community from their cages. We found that while the gavage did have long-lasting effects on the recipient animals that appeared to influence the inflammation phenotype, it did not eliminate either cage effects or the succession of the gut microbial community over time. Our results show that whether or not an initial gavage is used, experimental design must explicitly account for successional patterns over time and cage microenvironment or risk mis-interpretation that could lead to difficult to reproduce conclusions.

3.3 Results

3.3.1 Structuring The Microbial Community In A Study Of WT And Il10$^{-\backslash-}$ Mice

In order to characterize cage effects and how the gut microbial community changes over time in the presence and absence of inflammation, we collected repeated longitudinal samples from *Il10*$^{-/-}$ and WT mice (TABLE 3.1) at 2 weeks, 12 weeks and 20 weeks after removal from germ free conditions. In these experiments, mice were allowed to acquire the microbial community from their cage microenvironment. DNA was isolated from these samples and subjected to PCR targeting the V6 region of the 16S rRNA gene. Amplicons from these PCR reactions were characterized with paired-end HiSeq Illumina sequencing. Because our read length (100 bp) was longer than our amplicon size (75 bp), we had 2X coverage on every read and could therefore remove sequences if the paired sequences were not in good concordance (see methods). Merging the paired-end reads resulted in 55,153,918 consensus sequences that met all QA/QC criteria with an average length of 74.52 ± 1.05 (mean ± SD) (TABLE 3.2). Sequences were clustered into Operational Taxonomic Units (OTU) with the program AbundantOTU (see methods) and the consensus sequences from each OTU were matched to full length sequences in the Silva database, which were classified with the Ribosomal Database Project (RDP) classifier [51].

Just by visual inspection of assignments to the phyla level (FIGURE 3.1), we note (i) Proteobacteria are clearly higher in *Il10*$^{-/-}$ mice than in WT mice at early time points and (ii) in both *Il10*$^{-/-}$ and WT mice there is a shift in community structure over time from an early Firmicutes dominated community with more Proteobacteria to a community more dominated by Firmictes and Bacteriodetes with fewer Proteobacteria. This change in the composition of the microbial community is associated with a dramatic increase in richness for WT and a much less pronounced increase in richness for *Il10*$^{-/-}$mice (FIGURE 3.2A).

In order to explicitly consider the effects of cage, genotype and time, we performed PCoA using Bray-Curtis dissimilarity at the OTU level (FIGURE 3.3). From this analysis, time is clearly the dominant force in structuring the microbial community with a dramatic shift in both WT and *Il10*$^{-/-}$ animals from 2 weeks to 12 weeks and a smaller shift from 12 weeks to 20 weeks (FIGURE 3.3, top panel; see also supplementary FIGURE 3.S1 for a similar analysis with a different distance metric). Within the time shift, genotype remains an important factor with a noticeable separation between WT and *Il10*$^{-/-}$ mice at 2 weeks becoming essentially perfect separation at 20 weeks (FIGURE 3.3, middle panel).

From these results, one might conclude that the effects of genotype are substantial and become more pronounced over time. However, we used a nested experimental design in which WT and *Il10*$^{-/-}$ mice were housed in separate cages. If we look at the PCoA broken down by time-points and colored by cages (FIGURE 3.3, bottom panel), we see that the cage has a profound effect on the microbiome as there is strong clustering by cage.

To quantify the cage effect, we performed one-way ANOVAs with a factor of cage. Fitting these models for each cohort at each time point (FIGURE 3.4A and B) shows that cage effects can be pronounced. For nearly all of the first 5 principle co-ordinates in both WT and *Il10*$^{-/-}$ animals, we are able to reject a null hypothesis that cage has no effect at $p < 0.05$. We note, however, that the magnitude of cage effects can vary substantially over time. It appears in general that the 20 week time period for both WT and *Il10*$^{-/-}$ mice have more pronounced cage effects, but the development of the cage effects over time is noisy with the 2 week time point generally showing more pronounced cage effects than the 12 week timepoint for WT animals.

Given these pronounced cage effects, statistics that ignore cage will be in violation of the assumption of independence and will therefore produce faulty inference. In order to account for the effects of cage, genotype and time, therefore, we evaluated a mixed linear model in which genotype and time is a fixed effect and cage is a random effect (see methods). The form of this mixed linear model for the *Il10*$^{-/-}$ vs. WT experiments is:

$$Y = \text{genotype} + \text{time} + \text{genotype} \times \text{time} + (\text{genotype:cage}) + \text{error}$$

where genotype and time are fixed factors, genotype x time is a fixed factor interaction term and (genotype:cage) represents cage nested within genotype as a random effect, and Y is either PcOA axis value (supplemental FIGURE 3.S2; TABLE 3.S1), phylum count (supplemental TABLE 3.S2) or genus count (supplemental TABLE 3.S3) as called by the RDP classifier on a full-length reference sequences that matches our V6 tag sequences (see methods) or richness value (supplemental TABLE 3.S4). From this model we note that (i) time effects, genotype effects and time x genotype interactions are all highly significant for the first few principle co-ordiantes (FIGURE 3.S2); (ii) richness effects for time and genotype (FIGURE 3.2A) are highly significant (supplemental TABLE 3.S4). (iii) all phyla detected in our WT vs. *Il10*$^{-/-}$ experiment (FIGURE 3.1) changed significantly (FDR-corrected p-values <0.005) over time (supplementary TABLE 3.S2). However, only Proteobacteria and Verrucomicrobia had consistent genotype effects that were significant (with an FDR-corrected p-value of <0.10) independent of time. (iv) at the genus level (supplemental FIGURE 3.S5; TABLE 3.S3), the first 25 most significant effects are all due to time and the first 40 most significant effects are due to either time or genotype x time interactions. However, at a 10% FDR, 15 genera are significantly different due to genotype without a time effect, including the genera Escherichia_Shigella that contains *E. coli*.

Our results show that there are taxa that are associated with differences between the WT and *Il10*$^{-/-}$ genotypes independent of time. However, time and time x genotype interactions appear to explain more of the variance of the microbial community seen in our experiment. That is, effects associated with time and the interaction of time with genotype are much larger than effects associated with genotype, but there are some taxa that are different between WT and *Il10*$^{-/-}$ independent of time.

3.3.2 Gavage Can Modulate The Microbial Community

Because cage effects appeared so significant in our WT vs. *Il10*$^{-/-}$ experiment, we wished to better understand their cause. In the WT vs. *Il10*$^{-/-}$ experiment, the mice were removed from germ free conditions and therefore presumably had to acquire their microbiota from the cage micro-environment. We suspected that stochastic differences in the composition of the cages were driving the pronounced cage effects that we observed (FIGURE 3.4A -B). To test this hypothesis, we performed an additional experiment on WT mice in which a donor community generated from adult WT mice was given to one set of mice (hereafter referred to as the "gavage" treatment) while another set of mice were again allowed to acquire their microbial community from the cage environment (hereafter referred to as the "acquired" treatment). Fecal samples were collected at week 1, 2, 4 and 8 following removal from germ-free conditions and the microbal community was again characterized by Illumina sequencing targeting the V6 region of the 16S rRNA gene.

An examination of the results at the phyla level (FIGURE 3.5) demonstrates that at the 1 week time-point, the gavage treated animals appeared to have a microbial community that was in some ways a mixture of the donor community (FIGURE 3.5, upper right panel) and the community in the “acquired” group with a contribution from Proteobacteria (7.5%) in between the large fraction (41%) of Proteobacteria in acquired and the smaller fraction in the donor biota (1.5%). It appears, therefore, that the donor community influenced, but did not completely set, the resulting microbial community at 1 week. Over time, the fraction of Proteobacteria decreased in both the Acquired and Gavage groups. By week 8, the phyla view of these two groups was very similar (FIGURE 3.5, bottom panel). As was the case in the WT mice from our initial experiment, richness in both the gavage and treatment groups increased over time (FIGURE 3.2B) suggesting that the initial seeding of the microbial community by gavage did not give the gavage group a substantial “head start” in forming a mature microbial community.

To perform inference on the gavage experiment, we again used PcOA with Bray-Curtis as a distance metric for taxa clustered into OTUs (FIGURE 3.6; top panel). We see that, as in the WT vs. *Il10*$^{-/-}$ experiment, time is again a dominant force with clear separation of samples at different time points. However, at each time point gavage and acquired appear to be distinct (FIGURE 3.6; middle panel), although over the course of the experiment the differences between gavage and acquired become less pronounced (FIGURE 3.4C and D; FIGURE 3.6, top panel).

If initial differences in the microbial community drove cage effects, we might expect to see a reduced cage effect in the gavage group when compared to the acquired group. Examination of the PcOA plots colored by cage (FIGURE 3.6, bottom panel) appear to show pronounced cage effects in both the acquired and gavage groups, however, especially at later time points. To quantify this, we fit each treatment group at each time point with a one-way ANOVA with a fixed factor of cage. P-values generated from this model (FIGURE 3.4C and D, bottom panel) show that the gavage and acquired groups have a similar pattern of cage effects. At the 1 week time point (FIGURE 3.4C and D, bottom panel, black symbols), cage effects appear to be of marginal significance at best. At the 4 and 8 week timepoints, the cage effects have become much more pronounced for both gavage and acquired.

The presence of cage effects again suggests that a mixed linear model is an appropriate statistical framework in which to perform inference. To analyze the gavage vs. acquired dataset, we therefore formed the model:

Y = treatment + time + treatment x time + (treatment:cage) + error

where treatment is a fixed effect set to one value for gavaged animals and another for animals allowed to acquire the microbial community from the cages, time is a fixed factor, treatment x time is a fixed interaction term, and (treatment:cage) represents cage nested within treatment status as a random effect, Y is either PcOA axis value (supplemental FIGURE 3.S3; TABLE 3.S5), phylum count (supplemental TABLE 3.S6), genus count(supplemental TABLE 3.S7) or richness value (supplemental TABLE 3.S8).

From this model we conclude that time and time x treatment effects are generally more pronounced than the treatment effect confirming that the gavage effect did not eliminate the strong successional effect of time. Specifically, (i) richness changed over time but was not affected by treatment (supplementary TABLE 3.S8); (ii) at the phyla level (supplementary TABLE 3.S6) at a 10% FDR, time and treatment x time interactions are significant for all evaluated phyla; (iii)at the genus level, time and treatment x time represent the first 49 most significant effects (supplementary TABLE 3.S7).

3.3.3 Gavage Treatment Protects From Cage Effects Of Inflammation.

Even though gavage did not eliminate temporal effects, there were several taxa that were significantly different between gavage and acquired groups independent of time. At a 10% FDR cutoff at the phyla level (supplementary TABLE 3.S6), the treatment variable for both Bacteroidetes and Firmicutes were significantly different between gavage and acquired independent of time. At the genus level, the 10% FDR cutoff yielded 20 genera that were significantly different between gavage and acquired independent of the time factor (supplementary TABLE 3.S7). This suggests that, despite the overall progress on acquired and gavage to become more similar to each other (FIGURE 3.5), the gavage treatment did have some long lasting effects.

To study the consequences of these effects for host health, at the end of the gavage experiment, we sacrificed and scored a subset of the mice for inflammation. Inflammation was induced through the use of DSS (Dextran Sulfate Sodium) at the 8-week time point in the gavage study and after 12 days inflammation scores were established by histological examination of tissue. The mice that were gavaged did not show a cage effect for inflammation but the mice that were allowed to acquire the microbiome from the external environment showed distinct patterns of inflammation by cage (FIGURE 3.7). Interestingly, Lactobacillus, a taxa that is considered to have anti-inflammatory properties [83-85] was found to be significantly associated with time, treatment x time and treatment under the mixed linear model analysis of the gavage vs. acquired dataset (supplemental TABLE 3.S7). With our small sample size, we cannot meaningfully speculate on whether these associations are robust and would be reproducible in future cohorts. We also do not have sufficient data to know whether the gavage treatment in general will nullify the phenotypical influence local cage environments may contribute to experiments utilizing the DSS mouse model. Nonetheless, these data are intriguing in suggesting that long-term effects of an initial gavage may insulate an animal from environmentally-induced susceptibility to cage effects in phenotypes of interest.

3.3.4 Sequencing Depth And Quantitfying The Microbial Community

We have previously argued [20] that *E. coil* is one organism that may drive progression from inflammation to cancer. To verify the abundance of this organism, we performed qPCR targeting the 16S rRNA region at week 20 from the WT/ *Il10*$^{-/-}$ comparison (the qPCR data have also been previously published in Arthur et al.[20]). In order to validate our sequence data, we compared the qPCR data to the abundance of OTU 23, the OTU with the consensus sequence that most closely matches *E. coli* (FIGURE 3.8). We see a very strong quantitative relationship (FIGURE 3.8, top panel) validating both our sequence-based and qPCR quantification. In our experiment, we used an Illumina Hi-Seq pipeline that produced on average 450,000 sequences per sample that met our QA/QC criteria. We were curious to estimate what we would have seen if we had instead used technology such as 454 sequencing or Illumina MiSeq, which tends to produce approximately two orders of magnitude fewer sequences per sample. We therefore, sub-sampled our dataset and compared to the qPCR data (FIGURE 3.8, bottom panels). We see that much of the correlation with the qPCR data would have been lost if we had produced on average 4,500 sequences per sample (FIGURE 3.8, panel marked 1%). Consensus 23, the 23rd most abundant taxa in our dataset, represented ~1% of all of our sequences, but may still be biologically crucial. We conclude that it is not "overkill" to produce on the order of a million sequence per sample if there is interest in quantitative changes in the less abundant members of the microbial community and that such a sequence depth may be crucial in generating an accurate description of the microbial community.

3.4 Discussion

In our previous study [20] exploring the interactions of inflammation, tumor formation and the microbiota, we noticed a strong cage effect in which animals within the same cage had similar microbial communities. To explore the causes and consequences of these cage effects, we here used Illumina sequencing to characterize the 16S gene from fecal samples collected over time for four cohorts of mice following removal from germ-free conditions. In all of our cohorts, time was the dominant factor in structuring the microbial community with a low richness community with a substantial fraction of Proteobacteria becoming more like the adult mammalian gut over time with an increase in richness (FIGURE 3.2) and increasing domination by Bacteroidetes and Firmicutes (FIGURE 3.1, FIGURE 3.5). These results are similar to those seen in a successional sequencing experiment performed by [86]. Superimposed on these broad and reproducible successional patterns, however, our experiments also observed substantial individual variation in the microbial community that could in large part be explained by the cage in which the animals were housed (FIGURE 3.4). Starting the microbial community with an initial gavage from a mature gut microbial community influenced, but did not eliminate, the dependency on time or cage effects. Our data show that host genotype (WT vs. *Il10*$^{-/-}$), initial composition of the microbial community (gavage vs. acquired), selection pressure over time common across all cages and stochastic effects that develop differently over time in different cages all make important and measurable contributions in structuring the microbial community.

Protoebacteria contains many harmful pathogens including those that have been associated with gut inflammation, IBD and colorectal cancer [87]. In our "two hit" model [20] of cancer formation in *Il10*$^{-/-}$ mice, we argued that inflammation allows the Proteobacteria *E. coli* to invade the gut, where its genotoxic potential helps to drive progression to tumors in the *Il10*$^{-/-}$ mouse model. In the current experiments, Proteobacteria was one of two phyla (supplementary TABLE 3.S2) that were significantly different in *Il10*$^{-/-}$ mice then in WT independent of time while genus Escherichia.Shigella was one of 15 genera (supplementary TABLE 3.S3) that were different in *Il10*$^{-/-}$ vs. WT independent of time. This observation, that *E. coli* invades early in succession and is significantly higher independent of the other changes that occur in the microbial community over time is consistent with our two hit model in that it suggests that *E. coli*'s genotoxic influence on the host may begin well before the symptoms of inflammation become apparent in the *Il10*$^{-/-}$ mouse model.

Richness for all of our cohorts except for the *Il10*$^{-/-}$ group substantially increased over time (FIGURE 3.2). A plausible mechanism that would explain this observation is that the immune response and inflammation associated with the *Il10*$^{-/-}$ genotype prevents some taxa that are able to colonize WT animals from successfully becoming established within the inflamed *Il10*$^{-/-}$ gut. Our results are therefore consistent with the many [88-92] experiments that have shown a lower diversity in the human gut microbiota in IDB patients. Interestingly, while richness increased over time in our gavage and acquired experiment (FIGURE 3.2B), there was no significant difference in richness induced by the gavage treatment (p=0.42; Supplementary TABLE 3.S8). This suggests that even though the gavage treated mice were presumably exposed to a higher abundance of microbes than the acquired treatment group, this did not influence the number of microbes that ultimately successfully colonized the gavaged group.

In addition to using 16S sequencing to characterize the microbial community, we studied the functional consequences of cage effects by inducing inflammation through the use of DSS (Dextran Sulfate Sodium) on WT animals at 8-weeks after removal from germ free conditions (FIGURE 3.7). Inflammation scores established by histological examination of gut tissue showed that animals that were allowed to acquire their microbiota from the cage environment displayed a more pronounced cage effect in the degree of inflammation observed than animals whose microbial community was acquired by gavage. In our analysis utilizing mixed linear models with cage modeled as a random variable, treatment (gavage vs. acquired) effects on the structure of the microbial community were generally much less pronounced than effects due to time or interactions between time and treatment (supplementary TABLES 3.S5-3.S8). There were, however, 2 phyla (Bacteroidetes and Firmicutes; supplementary TABLE 3.S6) and 20 genera (including Lactobacillus) that were significantly different between gavage and acquired independently of time. Taken together, these observations suggest that an initial gavage can have subtle, long-lasting effects on the microbial community that have long-term consequences for host phenotype even as the microbial community changes substantially over time in ways that are different in different cages.

There is considerable evidence for cage effects in the literature. Microbial transfer of disease can be accomplished by housing WT mice with mice that have colitis [93] suggesting that within a cage, the microbial community can be shared between mice. Mice are known to eat feces and this presumably has a substantial effect on the microbial community. In this paper, we explicitly tested the hypothesis that cage effects are caused by initial differences in the microbial community within each cage. If these stochastic “founder” effects drive cage effects we would expect (i) cage effects to be pronounced at early time points and (ii) gavage to significantly mitigate cage effects. Neither of these predications were well supported by our data.In the gavage vs. acquired experiment for which we have the most temporal resolution at early time points, cage effects clearly become more pronounced over time (FIGURE 3.4C and D) moving from barely significant at week 1 to highly significant at weeks 4-8. Moreover, this same temporal pattern is seen in both the gavage and acquired groups. Our data, therefore, argue that “drift”, stochastic differences in the way the microbial community forms that are different in different cages, rather than founder effects, are the primary drivers of cage effects. Attempts to eliminate cage effects by standardizing the initial microbial community within cages or with identical initial gavage to multiple animals are therefore likely to fail.

The confounding nature of micro-environments on succession of microbial communities is a known problem in experimental design [86]. An outstanding recent paper [94] has argued that family transmission, if not properly accounted for, can lead to confounded experimental design and incorrect inference regarding the effects of genotype differences. We note that since our animals were born in germ-free conditions, family transmission is not a variable that can be considered in our experiments. However, by necessity, animals that have a similar path of family transmission have also shared cages. In animals not born under germ-free conditions, therefore, cage effects and family transmission effects are likely often confounded and this can be a further complication in experimental design.

In order to explicitly model the effects of cage, we utilized mixed linear models in which cage is set as a random effect. Mixed linear models have many advantages including a solid theoretical base [95, 96], wide utilization in the literature [97-100] and robust implementations in statistical packages such as R and SAS. However, mixed linear models impose an additional set of parametric assumptions over canonical linear models. In our case, they assume that the effects of cages are normally distributed with a mean of zero. These assumptions may be particularly inappropriate for metagenomics data, where it has been argued for the gut microbiome that only a few possible outcomes (or enterotypes) are likely [101]. While the enterotype hypothesis has been highly controversial [9, 25, 102], it seems unlikely that the opposite assumption that there is no repeatable structure to the microbial community within cages is unlikely to be broadly true. A finite subset of possible cage outcomes might therefore violate the assumptions of mixed linear models. With this in mind, we compared our results to a simple model in which the median value for each cage were fed into a canonical two way ANOVA (data not shown). We saw a broadly similar pattern of p-values with this approach, although as we might expect this median based linear models appeared to have substantially lower power than the full mixed linear model. Future research will undoubtedly pursue the question of the most appropriate model for cage effects for metagenomics experiments that makes the fewest assumptions while preserving the most power, but the broad concurrency of the median and mixed linear models is encouraging in that is suggests that our results are not primarily driven by the additional parametric assumptions about cage distribution in the mixed linear model.

It has long been a question in ecology how much community structure is driven by selection vs. stochastic events. In our data, we find evidence for both kinds of processes. The replicable drive to an end-point of a community dominated by Firmicutes and Bacteriodites (FIGURE 3.1, FIGURE 3.5) suggests selection pressure working over time in a reproducible matter to shape the gut microbial community. The fact that this pattern has been observed in other cohorts in the literature suggests the strength of these reproducible forces of selection [86]. Starting mice at the end point of succession by using gavage to introduce a mature gut community influenced, but did not eliminate, this stereotyped succession, which led to our surprising result thatmice treated with gavage looked more like donor mice at the end of our experiment rather than the beginning (FIGURE 3.5). Superimposed on this successional pattern, however, were cage effects (FIGURE 3.4C and D) which appear to develop over the first few weeks of removal from germ free conditions. We conclude that generation of robust and reproducible results for mouse models of the gut microbiota are dependent on explicit consideration of this variation in the microbial community induced by time and the cage micro-environment. Failure to consider these factors in both experimental design and statistical models is likely to lead to misinterpretation of experimental results in which changes due to cage or time are mistaken for the intended changes due to experimental manipulations.

3.5 Material And Methods

3.5.1 Illumina Sequence Pipeline

Following the protocol outlined by Arthur et. al. [20] we aligned the paired-end reads and merged overlapping sections to convert the paired reads into a single consensus sequence. Custom Java code was written in which the criteria for merging paired-end reads was two fold (i) an exact match to both the 5' and 3' primers was required and (ii) the overlapping region met or exceeded a 70 base threshold allowing for mismatches but not gaps. All nucleotides in common were selected for the consensus sequence; in cases where a base position had a disagreement between the two reads the base with the highest quality score was selected for inclusion. When an N was encountered in one read the base at the corresponding position in the read pair was selected. Consensus sequences were then used for clustering into Opperational Taxonomic Units (OTUs) by feeding them into AbundantOTU (http://omics.informatics.indiana.edu/AbundantOTU/) running on a linux box with 128 GB of RAM. AbundantOTU explicitly ignores rare taxa because of the higher propensity of error prone reads associated with the rare biosphere. OTUs generated in the form of consensus sequences by AbundantOTU were checked for potential chimeras using UCHIME (http://www.drive5.com/uchime/)[46] and the Gold reference database. For taxonomic classification the AbundantOTU consensus sequences were mapped to the Silva 108 database (http://www.arb-silva.de/) by BLAST with and e-score threshold of e-10. The top hits were selected and sent through RDP classifier version 2.1 (http://sourceforge.net/projects/rdp-classifier/) [51] with an RDP confidence threshold of 80% or greater used for assignment.

3.5.2 Statistical Analysis

OTU consensus sequences were collapsed into pivot table format where each row represents a sample and each column contains the raw counts for each OTU consensus sequence. Raw counts were transformed using a log frequency calculation (EQUATION 3.1) before use in downstream analysis.

$$Log_{10}\left(\frac{RC}{n} * \frac{\sum x}{N} + 1\right)$$

EQUATION 3.1: Log frequency normalization used to normalize raw OTU counts. Where RC represents the number of raw counts in a column cell (OTU, phyla, etc…) for a sample, n is the number of sequences in a sample, the sum of x is the total number of counts in the table and N is the total number of samples.

Bray-Curtis dissimilarity matrixes were generated from normalized data and Principle Co-ordinate Analysis (PCoA) was conducted through the use of the software package mothur [70]. We chose to use Bray-Curtis dissimilarity matrixes because the results obtained using UniFrac distances [103] and the QIIME software package [69] were broadly similar (supplemental Figure S1). Adjusting for cage effects was done by incorporating mixed linear models (EQUATION 3.2) utilizing SAS (supplemental code TABLE 3.S9) where cages were the random effects and genotype or treatment were fixed. Benjamini-Hochberg method for False Discovery Rate (FDR) correction was used for multiple testing correction. Diversity measurements of richness were calculated using custom Java code and rarefying down to the lowest number of sequences present in the data set. All statistical analysis was conducted either through R (http://www.r-project.org/), custom Java code (available upon request) and SAS (http://www.sas.com).

$$Y_{ijkl} = \mu + G_i + T_j + (GT)_{ij} + C_{k(i)} + \varepsilon_{ijkl}$$

EQUATION 3.2: Mixed effect linear models. Where Y_{ijkl} represents either PCoA axis value, phylum count, genus count or richness value for treatment/genotype *i*, time j, cage k and replicate l. G_i.is the effect of the i^{th} treatment/genotype. Treatment is set to one value for animals receiving gavage and another for animals allowed to acquire the microbial community from the cages, as genotype is set to either WT or *Il10*$^{-\backslash-}$. T_j is effect from the j^{th} time point. $(GT)_{ij}$ is the interaction effect between treatment/genotype i and time j. $C_{k(i)}$ is the effect from the k^{th} cage that is nested within the i^{th} treatment/genotype and ε_{ijkl} denotes the error associated with measuring Y_{ijkl}.

3.5.3 *Il10*$^{-\backslash-}$ Vs WT Study

Stool samples from germ-free WT 129/SvEv and *Il10*$^{-\backslash-}$ mice were collected at three time points (2 weeks, 12 weeks, 20 weeks). Due to an error in labeling one mouse (C5M5) was removed from each time point and not considered in any of the downstream analysis. Illumina sequencing of the 16S rRNA V6 hypervariable region generated a total of 55,153,918 reads that passed all QC steps which were ~75 base pairs in length (TABLE 3.2). Sequences were combined from 2 lanes of separate sequencing runs with lane 1 containing time points 2-weeks (20,990,204 reads) and 20-weeks (17,803,385 reads) and lane 2 contributing the 12-week time point (19,722,013 reads). AbundantOTU clustered the merged sequences into 1422 OTUs at a 97% threshold incorporating 99.993% of all sequences in a time of 616 minutes (~10 hours) and removing 375,807 singletons from downstream analysis (TABLE 3.2 and TABLE 3.5). Chimera detection using UCHIME identified 10 OTUs, which were then removed from the analysis pipeline.

Data from the 20 week time point has been previously published [20]. As noted in that paper, in both the WT and *Il10*$^{-\backslash-}$ mice, the carcinogen AOM was applied to mice in a subset of cages at week 4. Mice were housed 2-4 per cage in 6 WT and 5 *Il10*$^{-\backslash-}$ cages with WT cages (C4, C5, C6) and *Il10*$^{-\backslash-}$ cages (C1, C2, C3) receiving AOM treatment. As shown in our previous work (Figure 1 in Arthur et al [20]), the effects of AOM on the microbial community were much smaller than the effects of genotype. We did not therefore attempt to model the effects of AOM and did not separate AOM and non-AOM mice for the purposes of statistical inference. Since AOM cages were distinct from non-AOM cages, ignoring this variable may cause us to over-estimate the effects of cage at the 12 week and 20 week timepoints for the WT and *Il10*$^{-\backslash-}$. In general, however, the effects of AOM compared to the effects of cage were modest (data not shown), so we do not believe that the cage effects at these time points are being primarily driven by AOM induced differences between the cages. AOM was not applied in the gavage/acquired experiment, so the strong cage effects seen in this experiment (FIGURE 3.4C and D) cannot be explained by this potentially confounding variable.

3.5.4 Gavage Study

Sterile germ-free WT 129/SvEv mice were either inoculated by gavage from an amalgamation of 3 to 4 WT 129/SvEv donor fecal samples from mice ranging in age from 2 to 3 months or allowed to naturally acquire a microbiota. Stool samples were collected and processed for sequencing following the protocol outlined by Arthur et al [20] at the 1-week, 2-week, 4-week, and 8-week time points (TABLE 3.3). Mice were housed in 8 cages with 2-4 mice per cage (4 gavage cages and 4 acquired cages). One lane of paried-end Illumina 16S rRNA sequencing of the V6 hypervarialbe region produced 15,467,365 reads ~75 bases in length. The paired-end reads were merged into 14,880,760 sequences with an average length of 74.46 ± 1.18 (mean ± SD). Clustering by AbundantOTU took 106 minutes and produced 873 OTUs using a 97% threshold incorporating 99.996% of all sequences and removing 50,863 singletons from downstream analysis (TABLE 3.4 and TABLE 3.5). Chimera detection using UCHIME identified 5 OTUs, which we then removed from downstream analysis. At the 8-week time point the mice were exposed to DSS (Dextran Sulfate Sodium) in order to induce inflammation. After 12-days the mice were sacrificed and inflammation scores were cataloged through histological analysis of inflamed tissue.

3.6 Acknowledgements

We would like to thank Dr Timothy Tickle in the Department of Biostatistics at the Harvard School for Public Health, for his helpful suggestion regarding rarefaction visualization resulting in Figure 3.8.

3.7 Conflict Of Interest

Anthony Fodor has been a consultant to Salix.

3.8 Funding

This work was supported by NIH 2R01DK073338-05A1.

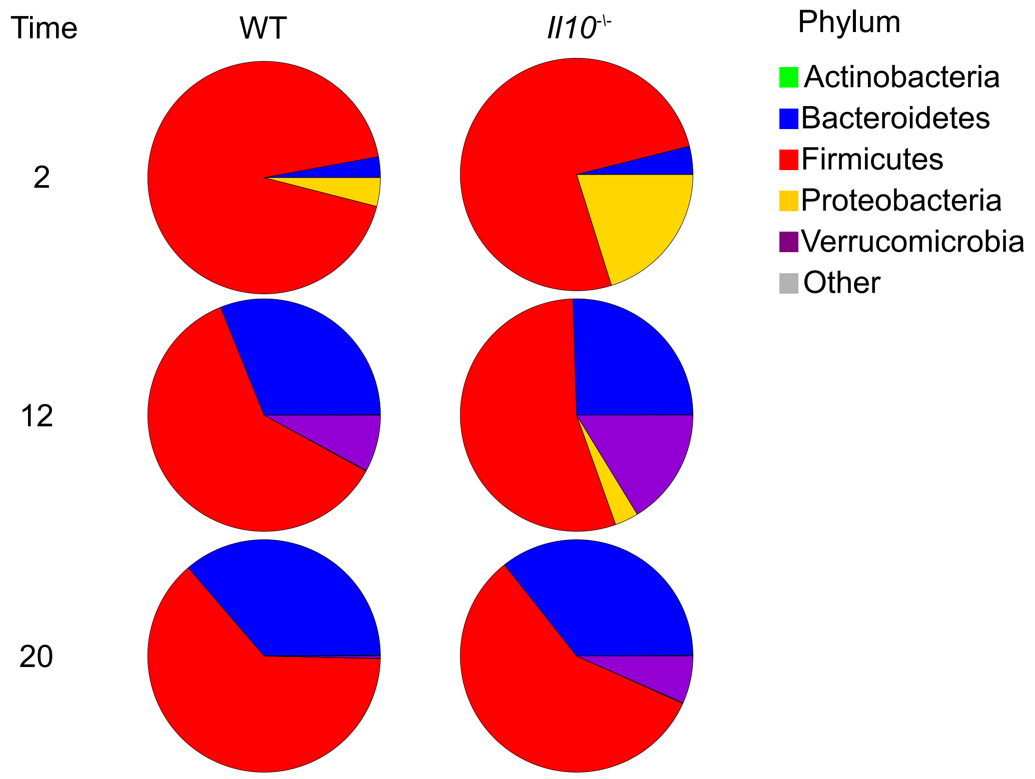


FIGURE 3.1: Alterations in microbial community composition over time at the phylum level in WT and *Il10*$^{-\setminus-}$ mouse model. Time is shown in the left column in weeks. Sample size at each time point included stool samples from 2 week: WT (n = 24) and *Il10*$^{-\setminus-}$ (n = 17), 12 week: WT (n = 22) and *Il10*$^{-\setminus-}$ (n = 16) and 20 week: WT (n = 24) and *Il10*$^{-\setminus-}$ (n = 15).

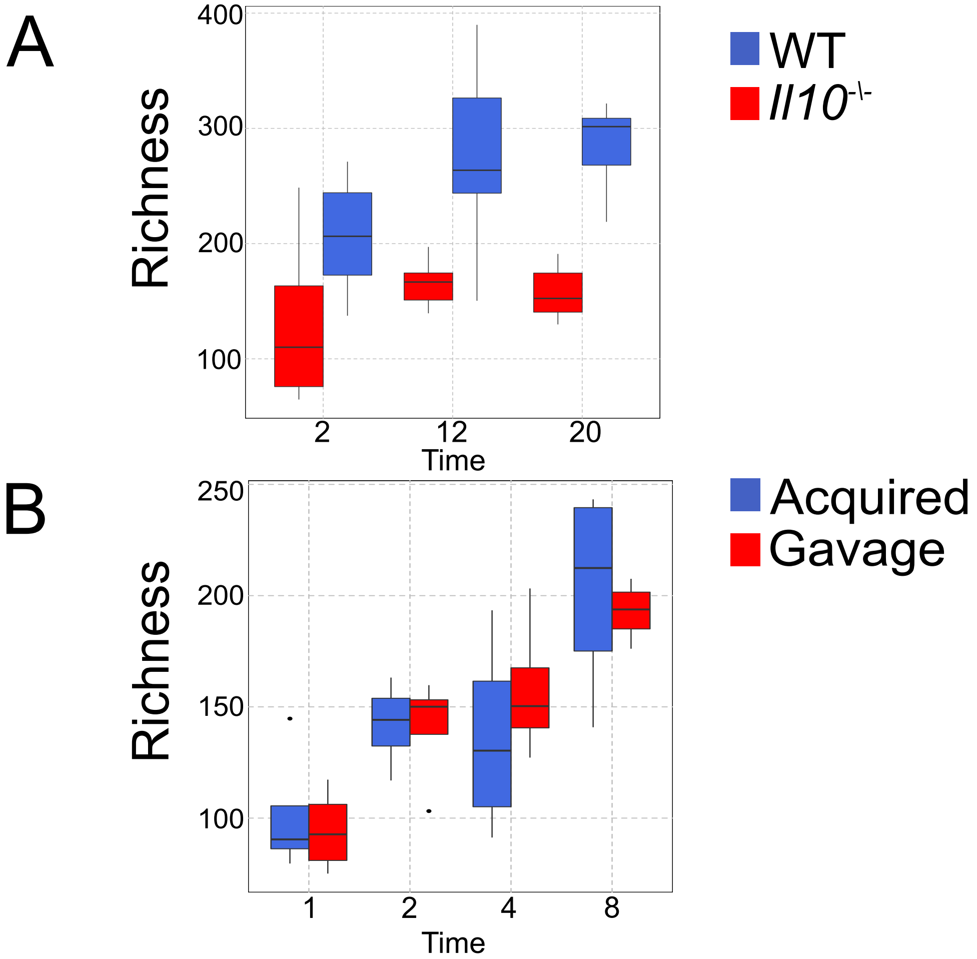


FIGURE 3.2: Richness as a function of time for the WT vs *Il10*$^{-\backslash-}$ experiment (A) and the acquired vs gavage experiment (B). The Donor biota for the gavage experiment has a richness value of 118.8, similar to the week 1 values. Richness was corrected for the number of sequences collected in each sample (see methods).

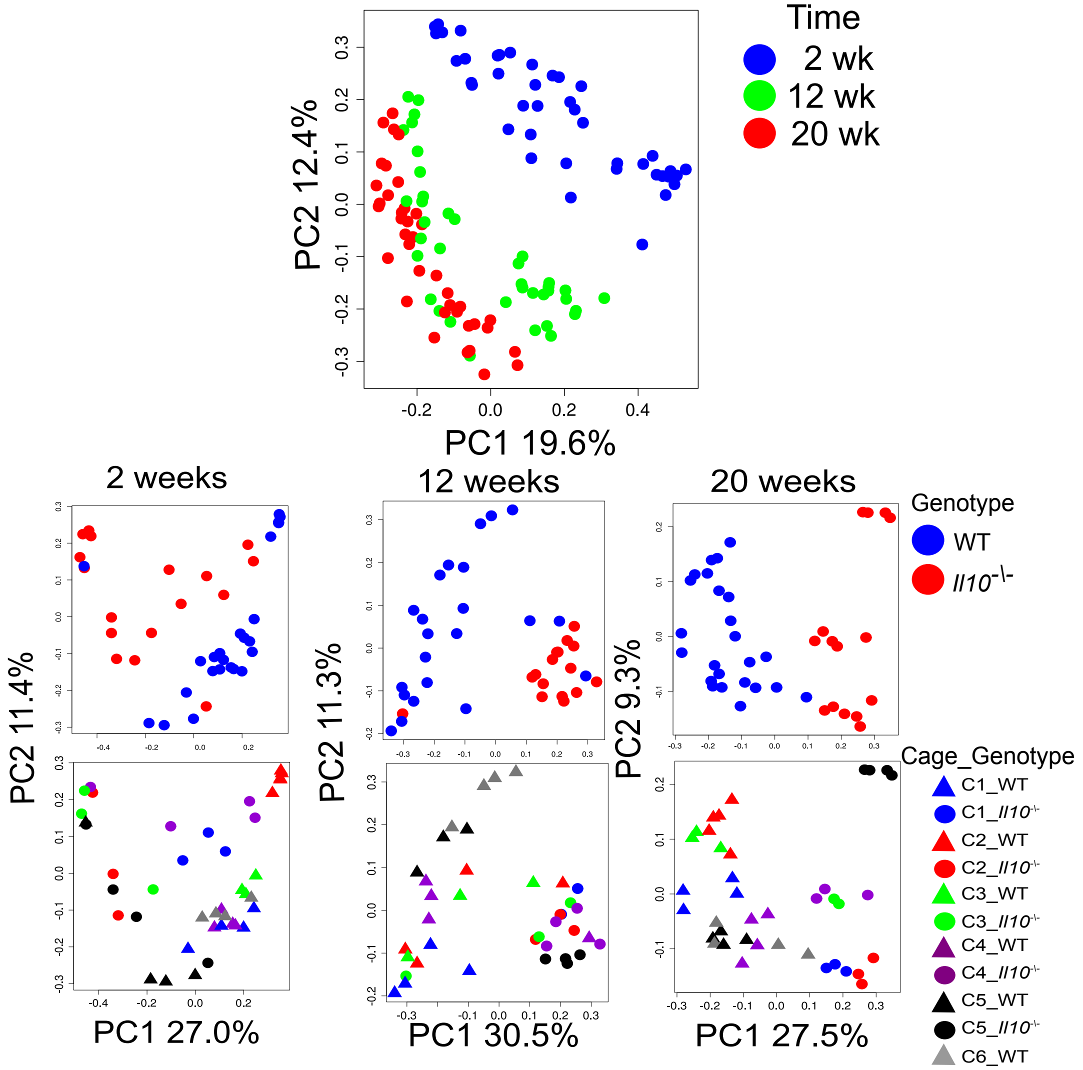


FIGURE 3.3: Bray-Curtis dissimilarity PCoA at the OTU level showing a clear separation between early and late time points (top panel). Independent PCoA clustering was performed for each time point and are colored by genotype (middle panel) and cage (bottom panel).

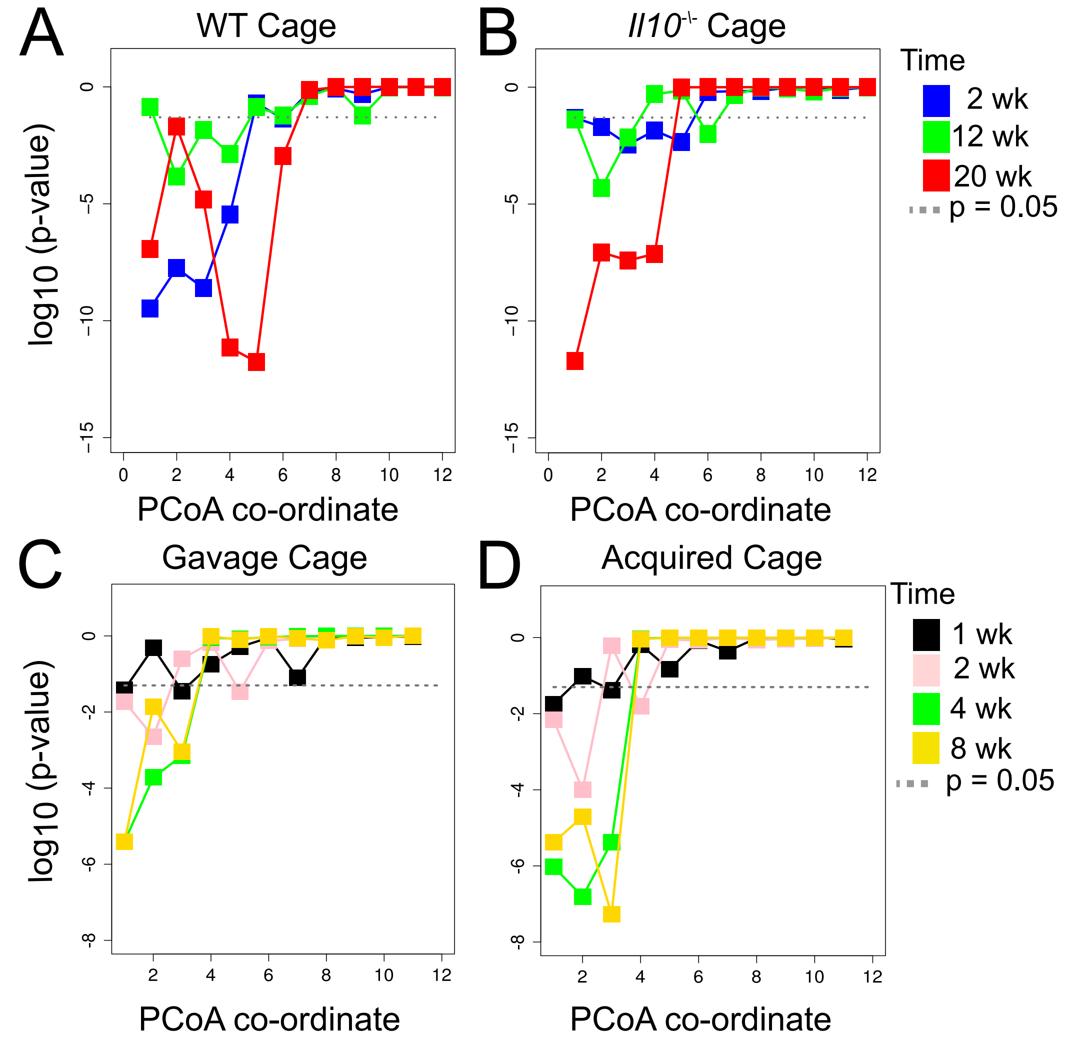


FIGURE 3.4: Cage effects illustrated through the use of Bray-Curtis PCoA performed at the OTU level. Shown for the first 12 PCoA co-ordinates are the p-values from a one-way ANOVA with a fixed factor of cage evaluating the null hypothesis that cage had no effect on the distribution of the co-ordinate.

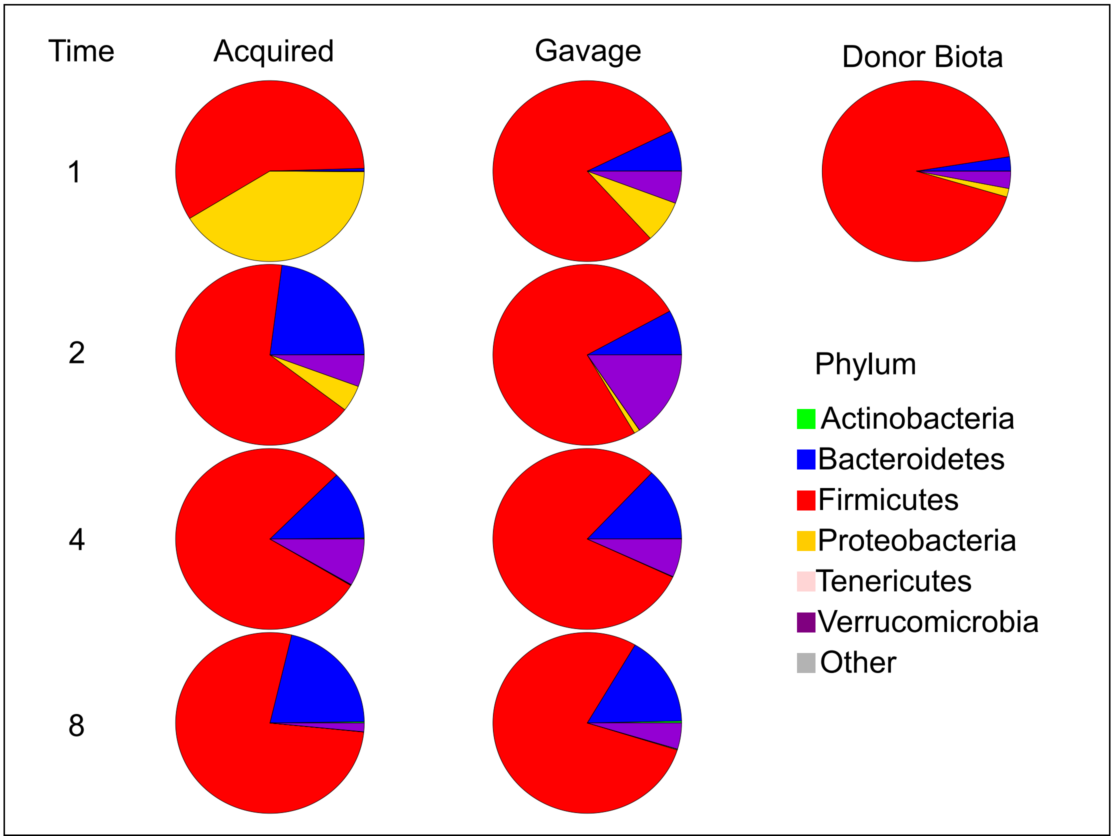


FIGURE 3.5: Alterations in microbial community composition at the phylum level over time in mice for which initial gavage was performed and mice allowed to acquire the microbial community from the cage environment. Time is shown in the left column in weeks. The microbial community composition for the donor is shown in the upper right-hand corner. Sample size at each time point was uniform with Gavage (n = 12) and Acquired (n = 12) with the Donor Biota having only 1 sample (n = 1).

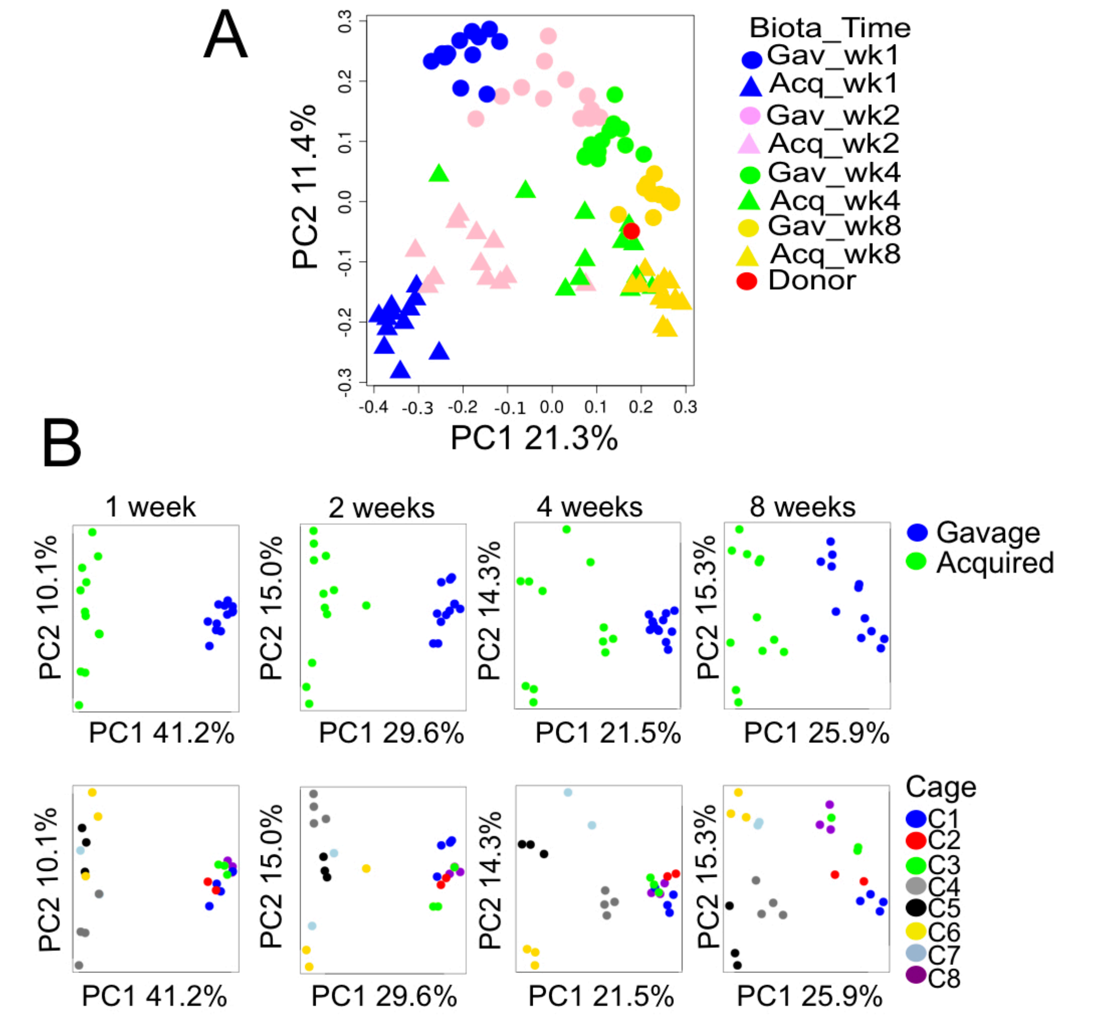


FIGURE 3.6: (A) Bray-Curtis dissimilarity PCoA at the OTU level showing microbial community shifting over time. (B) Independent PCoA clustering was performed for each time point and are colored by gavage status (top panel) and cage (bottom panel).

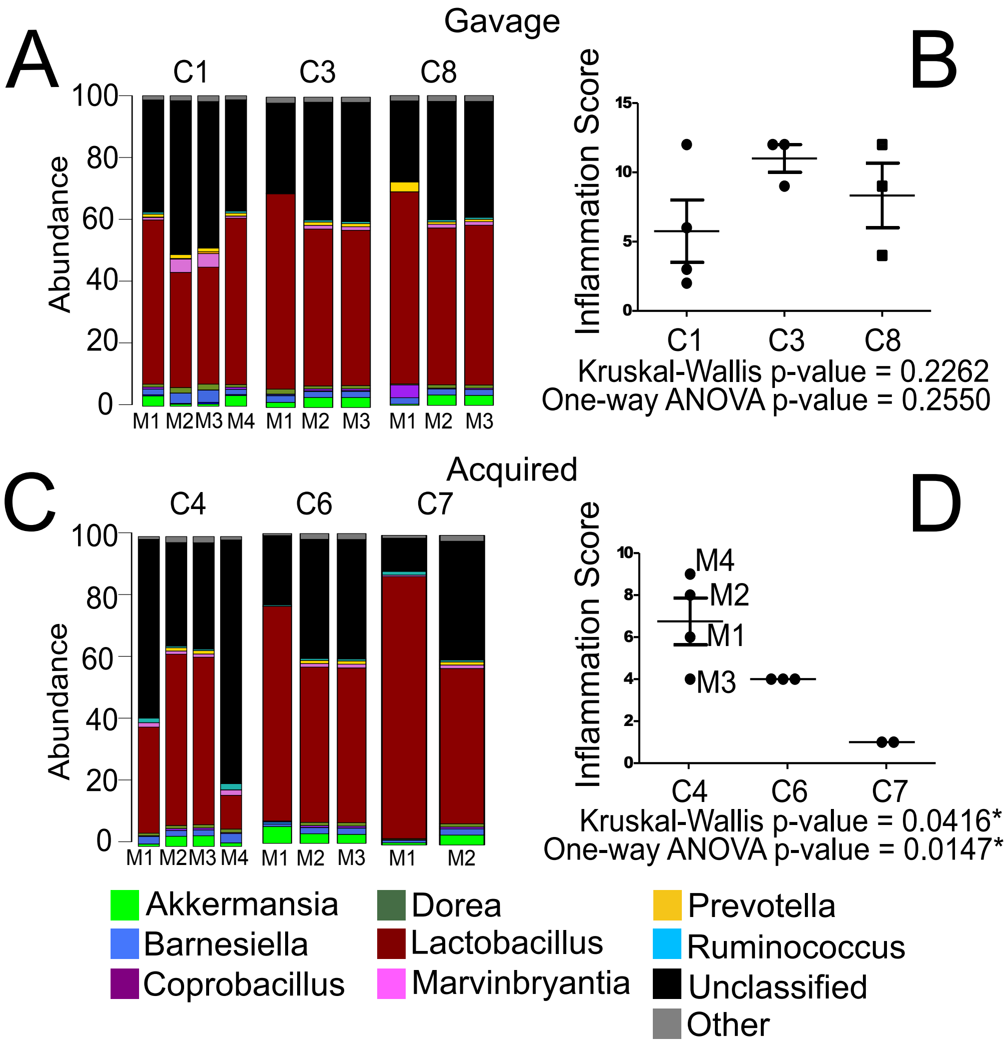


FIGURE 3.7: Relative abundance of genera in (A) Gavage and (C) Acquired at the 8 week time point broken down by cages. Each bar represents an individual mouse that received DSS. Differences in inflammation scores with a factor of cage were not significant for the gavage mice (B) but were for the Acquired mice (D) with both parametric One-Way ANOVA and a non-parametric Kruskal-Wallis (with the indicated p-values) *P < 0.05.

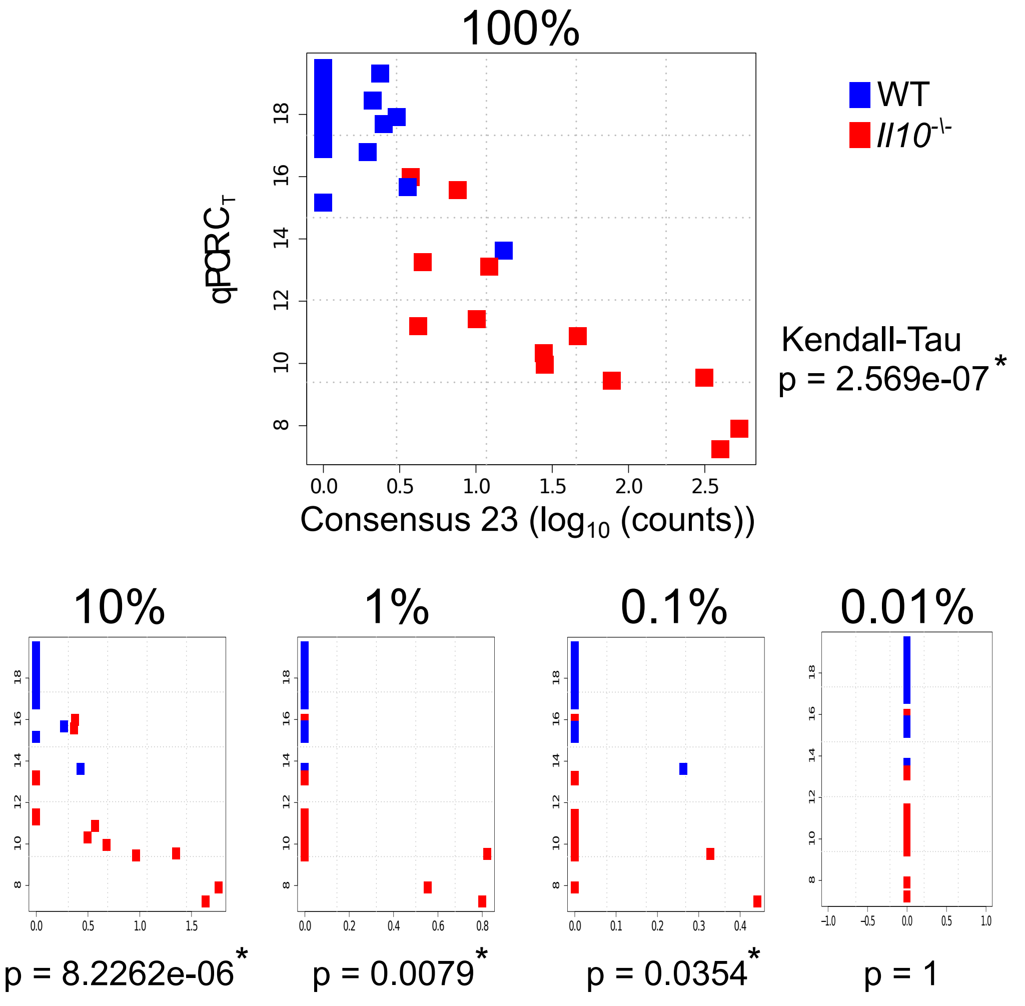


FIGURE 3.8: The ability to detect low abundant taxa is dependent upon sequencing depth. (Top Panel): C_T for qPCR for *E. coli* specific-C_T from 16S rRNA universal primers plotted against the V6 consensus sequence with the best match to *E. coli* 16S sequence. (Bottom Panel): Our samples had an average of ~450,000 V6 Illumina sequences per sample, but even a factor of 10 reduction would have impaired our ability to quantitatively detect *E. coli* even though *E. coli* was the 23rd most abundant taxa among all taxa. All p-values shown are from Kendall's tau. *P < 0.05.

TABLE 3.1: WT vs *Il10*$^{-\backslash-}$ study, number of stool samples used in downstream analysis. Removed mouse C5M5 because it was mislabeled. Mice were housed in 6 WT cages and 5 *Il10*$^{-\backslash-}$ cages.

Time Point (week)	**WT**		***Il10***$^{-\backslash-}$		**Total**
	NOAOM	**AOM+**	**NOAOM**	**AOM+**	
2	12	12	8	9	41
12	11	11	8	8	38
20	12	12	7	8	39
					118

TABLE 3.2: Number of reads in WT vs *Il10*$^{-\backslash-}$ study before and after QC. After QC a total of 3,3614,684 reads were removed and the remaining reads have a length of 74.52 ± 1.05.

Time	**Raw Reads**	**QC Reads**
2	20,990,204	19,112,305
12	19,722,013	19,127,448
20	17,803,385	16,914,165
Total	58,515,602	55,153,918

TABLE 3.3: Acquired vs Gavage study stool samples of WT mice. The Donor used for gavaging is an amalgamation of male and female WT mice ranging in age from 2 – 3 months. Mice were housed in 4 Gavage cages and 4 Acquired cages.

Time Point (week)	**Acquired**	**Gavage**	**Total**
1	12	12	24
2	12	12	24
4	12	12	24
8	12	12	24
			96

TABLE 3.4: Number of reads in Gavage Study before and after QC. After QC a total of 586,605 sequences were removed and the remaining reads have a length of 74.46 ± 1.18.

Time	**Raw Reads**	**QC Reads**
1	4,008,877	3,851,944
2	4,282,733	4,113,112
4	2,643,686	2,504,576
8	3,903,810	3,799,766
Donor	628,259	611,362
Total	15,467,365	14,880,760

TABLE 3.5: Initial reads were feed into AbundantOTU running on a linux box with 128 GB of RAM. Singletons identified by AbundantOTU were removed from downstream analysis.

Study	**Initial Reads**	**OTUs**	**Singletons**	**Run Time**
WT vs *Il10*$^{-\backslash-}$	55,153,918	1422	375,807	616 mins (10.3 hrs)
Gavage vs Acquired	14,880,760	873	50,863	106 mins (1.8 hrs)

TABLE 3.6: WT/*Il10*$^{-\backslash-}$ study. Richness rarified to 19,000 sequences for the Mixed Linear Model and Richness rarified to 19,000 sequences for the median cage value model. Both models show a significant time and genotype effect but no. P-values are corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure.
*P < 0.05.

Effect	**Mixed Linear Model**	**Median Cage Value**
Time	0.0016*	0.03*
Genotype	0.0016*	2.3e-05*
Genotype x Time	0.309	0.257

TABLE 3.7: Gavage study. Richness rarified to 2,000 sequences for the Mixed Linear Model and Richness rarified to 11,000 sequences for the median cage value model. Both models show a significant time effect. P-values are corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. *P < 0.05.

Effect	**Mixed Linear Model**	**Median Cage Value**
Time	5.48e-09*	7.101e-06*
Treatment	0.64	0.9332
Treatment x Time	0.98	1.3692

REFERENCES

1. Edwards, K.J., T.M. Gihring, and J.F. Banfield, *Seasonal variations in microbial populations and environmental conditions in an extreme acid mine drainage environment.* Applied and environmental microbiology, 1999. **65**(8): p. 3627-32.

2. Moreau, J.W., R.A. Zierenberg, and J.F. Banfield, *Diversity of dissimilatory sulfite reductase genes (dsrAB) in a salt marsh impacted by long-term acid mine drainage.* Applied and environmental microbiology, 2010. **76**(14): p. 4819-28.

3. von Wintzingerode, F., U.B. Gobel, and E. Stackebrandt, *Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis.* FEMS microbiology reviews, 1997. **21**(3): p. 213-29.

4. Banning, N.C., et al., *Soil microbial community successional patterns during forest ecosystem restoration.* Applied and environmental microbiology, 2011. **77**(17): p. 6158-64.

5. Roesch, L.F., et al., *Pyrosequencing enumerates and contrasts soil microbial diversity.* The ISME journal, 2007. **1**(4): p. 283-90.

6. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.

7. Fodor, A.A., et al., *The "most wanted" Taxa from the Human Microbiome for Whole Genome Sequencing.* PloS one, 2012. **7**(7): p. e41294.

8. Huttenhower, C., et al., *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207-214.

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

10. Peterson, J., et al., *The NIH Human Microbiome Project.* Genome research, 2009. **19**(12): p. 2317-23.

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

12. Plassart, P., et al., *Evaluation of the ISO standard 11063 DNA extraction procedure for assessing soil microbial abundance and community structure.* PloS one, 2012. **7**(9): p. e44279.

13. Zhao, J., et al., *Impact of enhanced Staphylococcus DNA extraction on microbial community measures in cystic fibrosis sputum.* PloS one, 2012. **7**(3): p. e33127.

14. HMP-C, *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-21.

15. Hayashi, H., M. Sakamoto, and Y. Benno, *Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods.* Microbiology and Immunology, 2002. **46**(8): p. 535-548.

16. Lane, D.J., et al., *Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses.* Proceedings of the National Academy of Sciences of the United States of America, 1985. **82**(20): p. 6955-9.

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

18. Spencer MD, H.T., Reid RW, Fischer LM, Zeisel SH, Fodor AA., *Association Between Composition of the Human Gastrointestinal Microbiome and Development of Fatty Liver With Choline Deficiency.* Gastroenterology, 2011. **140**(3): p. 976-986.

19. Gloor, G.B., et al., *Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products.* PloS one, 2010. **5**(10): p. e15406.

20. Arthur, J.C., et al., *Intestinal inflammation targets cancer-inducing activity of the microbiota.* Science, 2012. **338**(6103): p. 120-3.

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

22. Hamady, M., et al., *Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex.* Nature methods, 2008. **5**(3): p. 235-7.

23. Grice, E.A., et al., *Topographical and temporal diversity of the human skin microbiome.* Science, 2009. **324**(5931): p. 1190-2.

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

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

26. Fisher, M.M. and E.W. Triplett, *Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities.* Appl Environ Microbiol, 1999. **65**(10): p. 4630-6.

27. Liu, W.T., et al., *Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA.* Appl Environ Microbiol, 1997. **63**(11): p. 4516-22.

28. Fischer, S.G. and L.S. Lerman, *Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis.* Cell, 1979. **16**(1): p. 191-200.

29. McCafferty, J., et al., *Peak Studio: a tool for the visualization and analysis of fragment analysis files.* Environmental Microbiology Reports, 2012. **4**(5): p. 556-561.

30. Corrigan, A., et al., *Effect of dietary supplementation with a Saccharomyces cerevisiae mannan oligosaccharide on the bacterial community structure of broiler cecal contents.* Applied and environmental microbiology, 2011. **77**(18): p. 6653-62.

31. Or, A. and U. Gophna, *Detection of Spatial and Temporal Influences on Bacterial Communities in an Urban Stream by Automated Ribosomal Intergenic Ribosomal Spacer Analysis.* Microbes and environments / JSME, 2011. **26**(4): p. 360-366.

32. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(12): p. 5463-7.

33. MacLean, D., J.D. Jones, and D.J. Studholme, *Application of 'next-generation' sequencing technologies to microbial genetics.* Nat Rev Microbiol, 2009. **7**(4): p. 287-96.

34. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.

35. Petrosino, J.F., et al., *Metagenomic pyrosequencing and microbial identification.* Clin Chem, 2009. **55**(5): p. 856-66.

36. Shah, N., et al., *COMPARING BACTERIAL COMMUNITIES INFERRED FROM 16S rRNA GENE SEQUENCING AND SHOTGUN METAGENOMICS.* Pac Symp Biocomput, 2010: p. 165-76.

37. Sogin, M.L., et al., *Microbial diversity in the deep sea and the underexplored "rare biosphere".* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(32): p. 12115-20.

38. Hamp, T.J., W.J. Jones, and A.A. Fodor, *Effects of experimental choices and analysis noise on surveys of the "rare biosphere".* Applied and environmental microbiology, 2009. **75**(10): p. 3263-70.

39. Kim, M., M. Morrison, and Z. Yu, *Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes.* Journal of microbiological methods, 2011. **84**(1): p. 81-7.

40. Huse, S.M., et al., *Ironing out the wrinkles in the rare biosphere through improved OTU clustering.* Environ Microbiol, 2010. **12**(7): p. 1889-98.

41. Kunin, V., et al., *Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.* Environ Microbiol, 2010. **12**(1): p. 118-23.

42. Huse, S.M., et al., *Accuracy and quality of massively parallel DNA pyrosequencing.* Genome Biol, 2007. **8**(7): p. R143.

43. Minoche, A.E., J.C. Dohm, and H. Himmelbauer, *Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.* Genome Biol, 2011. **12**(11): p. R112.

44. Quince, C., et al., *Accurate determination of microbial diversity from 454 pyrosequencing data.* Nat Methods, 2009. **6**(9): p. 639-41.

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

46. Edgar, R.C., et al., *UCHIME improves sensitivity and speed of chimera detection.* Bioinformatics, 2011. **27**(16): p. 2194-200.

47. Hugenholtzt, P. and T. Huber, *Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases.* International journal of systematic and evolutionary microbiology, 2003. **53**(Pt 1): p. 289-93.

48. Ashelford, K.E., et al., *At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies.* Applied and environmental microbiology, 2005. **71**(12): p. 7724-36.

49. Haas, B.J., et al., *Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons.* Genome research, 2011. **21**(3): p. 494-504.

50. Liu, Z., et al., *Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers.* Nucleic Acids Res, 2008. **36**(18): p. e120.

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

52. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Appl Environ Microbiol, 2006. **72**(7): p. 5069-72.

53. Pruesse, E., et al., *SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.* Nucleic Acids Res, 2007. **35**(21): p. 7188-96.

54. McDonald, D., et al., *An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.* The ISME journal, 2012. **6**(3): p. 610-8.

55. Huse, S.M., et al., *Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing.* PLoS genetics, 2008. **4**(11): p. e1000255.

56. Schloss, P.D., *A high-throughput DNA sequence aligner for microbial ecology studies.* PloS one, 2009. **4**(12): p. e8230.

57. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(16): p. 9440-5.

58. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

59. Barrantes, G. and L. Sandoval, *Conceptual and statistical problems associated with the use of diversity indices in ecology.* Revista De Biologia Tropical, 2009. **57**(3): p. 451-460.

60. Ramette, A., *Multivariate analyses in microbial ecology.* FEMS microbiology ecology, 2007. **62**(2): p. 142-60.

61. Lozupone, C. and R. Knight, *UniFrac: a new phylogenetic method for comparing microbial communities.* Applied and environmental microbiology, 2005. **71**(12): p. 8228-35.

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

63. Knights, D., E.K. Costello, and R. Knight, *Supervised classification of human microbiota.* FEMS microbiology reviews, 2011. **35**(2): p. 343-59.

64. Aagaard, K., et al., *A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy.* PloS one, 2012. **7**(6): p. e36466.

65. Holmes, I., K. Harris, and C. Quince, *Dirichlet multinomial mixtures: generative models for microbial metagenomics.* PloS one, 2012. **7**(2): p. e30126.

66. Waldron, L., et al., *Optimized application of penalized regression methods to diverse genomic data.* Bioinformatics, 2011. **27**(24): p. 3399-406.

67. Liu, Z., et al., *Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data.* Bioinformatics, 2011. **27**(23): p. 3242-9.

68. Sboner, A., et al., *The real cost of sequencing: higher than you think!* Genome biology, 2011. **12**(8): p. 125.

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

70. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and environmental microbiology, 2009. **75**(23): p. 7537-41.

71. Giovannoni, S.J., et al., *Genetic Diversity in Sargasso Sea Bacterioplankton.* Nature, 1990. **345**(6270): p. 60-63.

72. Garcia-Martinez, J., et al., *Use of the 16S--23S ribosomal genes spacer region in studies of prokaryotic diversity.* J Microbiol Methods, 1999. **36**(1-2): p. 55-64.

73. Borneman, J. and E.W. Triplett, *Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation.* Appl Environ Microbiol, 1997. **63**(7): p. 2647-53.

74. Danovaro, R., et al., *Comparison of two fingerprinting techniques, terminal restriction fragment length polymorphism and automated ribosomal intergenic spacer analysis, for determination of bacterial diversity in aquatic environments.* Appl Environ Microbiol, 2006. **72**(9): p. 5982-9.

75. Culman, S.W., et al., *T-REX: software for the processing and analysis of T-RFLP data.* BMC Bioinformatics, 2009. **10**: p. 171.

76. Abdo, Z., et al., *Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes.* Environ Microbiol, 2006. **8**(5): p. 929-38.

77. Scallan, U., et al., *ribosort: a program for automated data preparation and exploratory analysis of microbial community fingerprints.* Molecular ecology resources, 2008. **8**(1): p. 95-8.

78. Ishii S, K.K., Senoo K., *Application of a clustering-based peak alignment algorithm to analyze various DNA fingerprinting data.* J Microbiol Methods, 2009. **78**(3): p. 344-50.

79. Andrade, L. and E.S. Manolakos, *Signal background estimation and baseline correction algorithms for accurate DNA sequencing.* Journal of Vlsi Signal Processing Systems for Signal Image and Video Technology, 2003. **35**(3): p. 229-243.

80. Savitzky, A., Golay, M. J. E. , *Smoothing and Differentiation of Data by Simplified Least Squares Procedures.* Anal. Chem., 1964. **36**(8): p. 1627-1639.

81. Ursell, L.K., et al., *The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites.* J Allergy Clin Immunol. **129**(5): p. 1204-8.

82. Knoch, B., et al., *Diversity of caecal bacteria is altered in interleukin-10 gene-deficient mice before and after colitis onset and when fed polyunsaturated fatty acids.* Microbiology-Sgm, 2010. **156**: p. 3306-3316.

83. Santos Rocha, C., et al., *Anti-inflammatory properties of dairy lactobacilli.* Inflammatory bowel diseases, 2012. **18**(4): p. 657-66.

84. Servin, A.L., *Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens.* FEMS microbiology reviews, 2004. **28**(4): p. 405-40.

85. von Schillde, M.A., et al., *Lactocepin secreted by Lactobacillus exerts anti-inflammatory effects by selectively degrading proinflammatory chemokines.* Cell host & microbe, 2012. **11**(4): p. 387-96.

86. Gillilland, M.G., 3rd, et al., *Ecological succession of bacterial communities during conventionalization of germ-free mice.* Applied and environmental microbiology, 2012. **78**(7): p. 2359-66.

87. Carvalho, F.A., et al., *Transient inability to manage proteobacteria promotes chronic gut inflammation in TLR5-deficient mice.* Cell host & microbe, 2012. **12**(2): p. 139-52.

88. Baumgart, M., et al., *Culture independent analysis of ileal mucosa reveals a selective increase in invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum.* The ISME journal, 2007. **1**(5): p. 403-18.

89. Frank, D.N., et al., *Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(34): p. 13780-5.

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

91. Ott, S.J., et al., *Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease.* Gut, 2004. **53**(5): p. 685-93.

92. Zhang, M., et al., *Structural shifts of mucosa-associated lactobacilli and Clostridium leptum subgroup in patients with ulcerative colitis.* Journal of clinical microbiology, 2007. **45**(2): p. 496-500.

93. Elinav, E., et al., *NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis.* Cell, 2011. **145**(5): p. 745-57.

94. Ubeda, C., et al., *Familial transmission rather than defective innate immunity shapes the distinct intestinal microbiota of TLR-deficient mice.* The Journal of experimental medicine, 2012. **209**(8): p. 1445-56.

95. Raudenbush, S.W. and A.S. Bryk, *Hierarchical linear models : applications and data analysis methods*. 2nd ed. Advanced quantitative techniques in the social sciences 12002, Thousand Oaks: Sage Publications. xxiv, 485 p.

96. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.

97. Brown, C.T., et al., *Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes.* PloS one, 2011. **6**(10): p. e25792.

98. Listgarten, J., et al., *Improved linear mixed models for genome-wide association studies.* Nat Methods. **9**(6): p. 525-6.

99. Vilhjalmsson, B.J. and M. Nordborg, *The nature of confounding in genome-wide association studies.* Nature reviews. Genetics.

100. Ross , E.M., er al., *High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing.* BMC Genetics, 2012. **13**(53).

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

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

103. Lozupone, C., et al., *UniFrac: an effective distance metric for microbial community comparison.* The ISME journal, 2011. **5**(2): p. 169-72.

APPENDIX A: SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Supplemental Figures

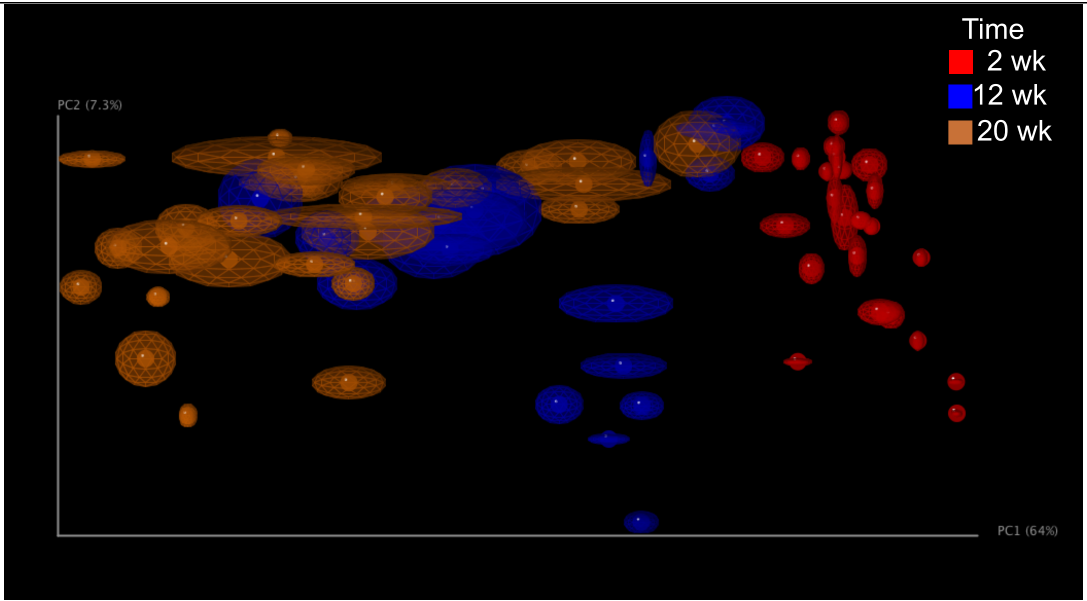


FIGURE 3.S1: WT vs *Il10*$^{-\backslash-}$ study. UniFrac distance metric produces the same pattern of clustering as that of a Bray-Curtis dissimilarity matrix. The above figure is the result of the default settings in Qiime and jackknife resampling of 110 sequences per sample. The orbs surrounding each data point are 95% confidence intervals based on the jackknife re-sampling.

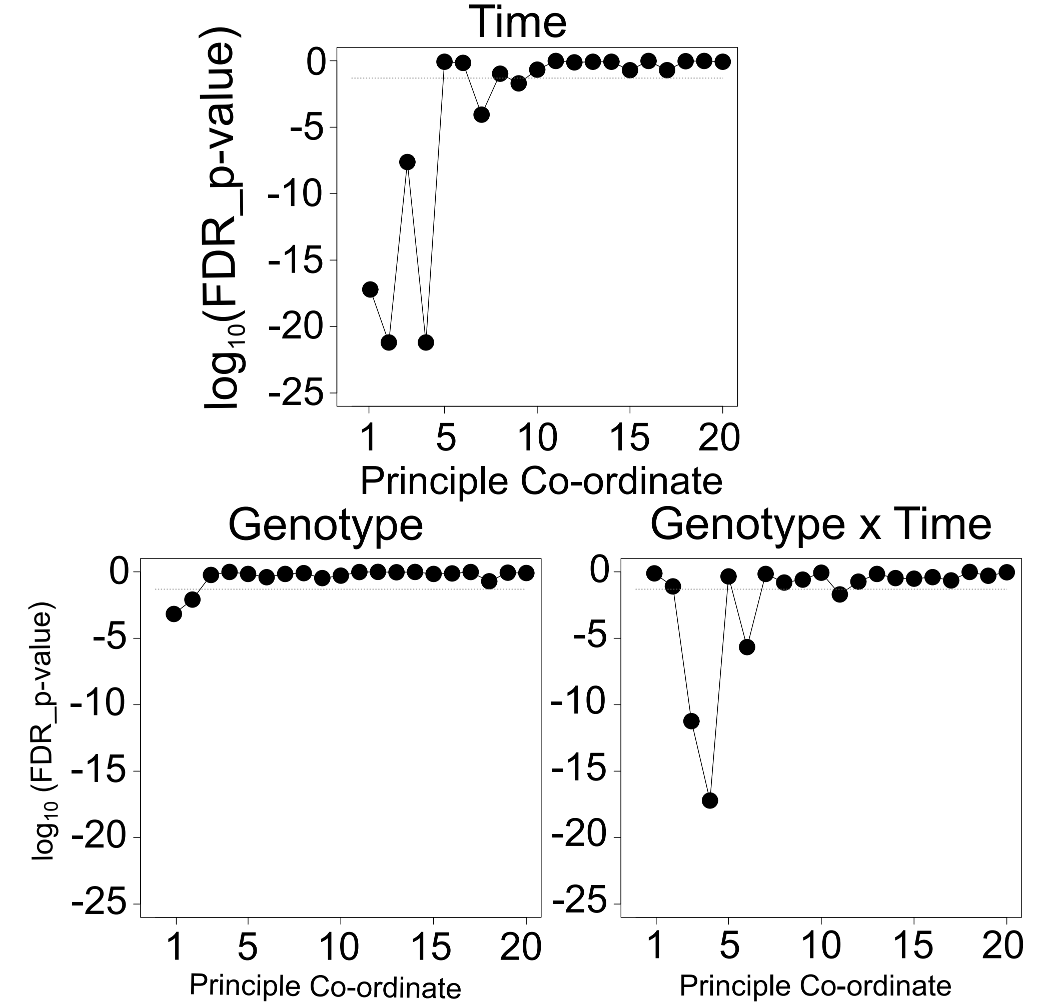


FIGURE 3.S2: WT vs *Il10*$^{-\backslash-}$ study. Results of a mixed effect linear model with PCoA co-ordinates as the dependent variable. The effects of Time, Genotype, and the Genotype x Time interaction show significance through the first few PCoA co-ordinates. Dotted line represents P = 0.05.

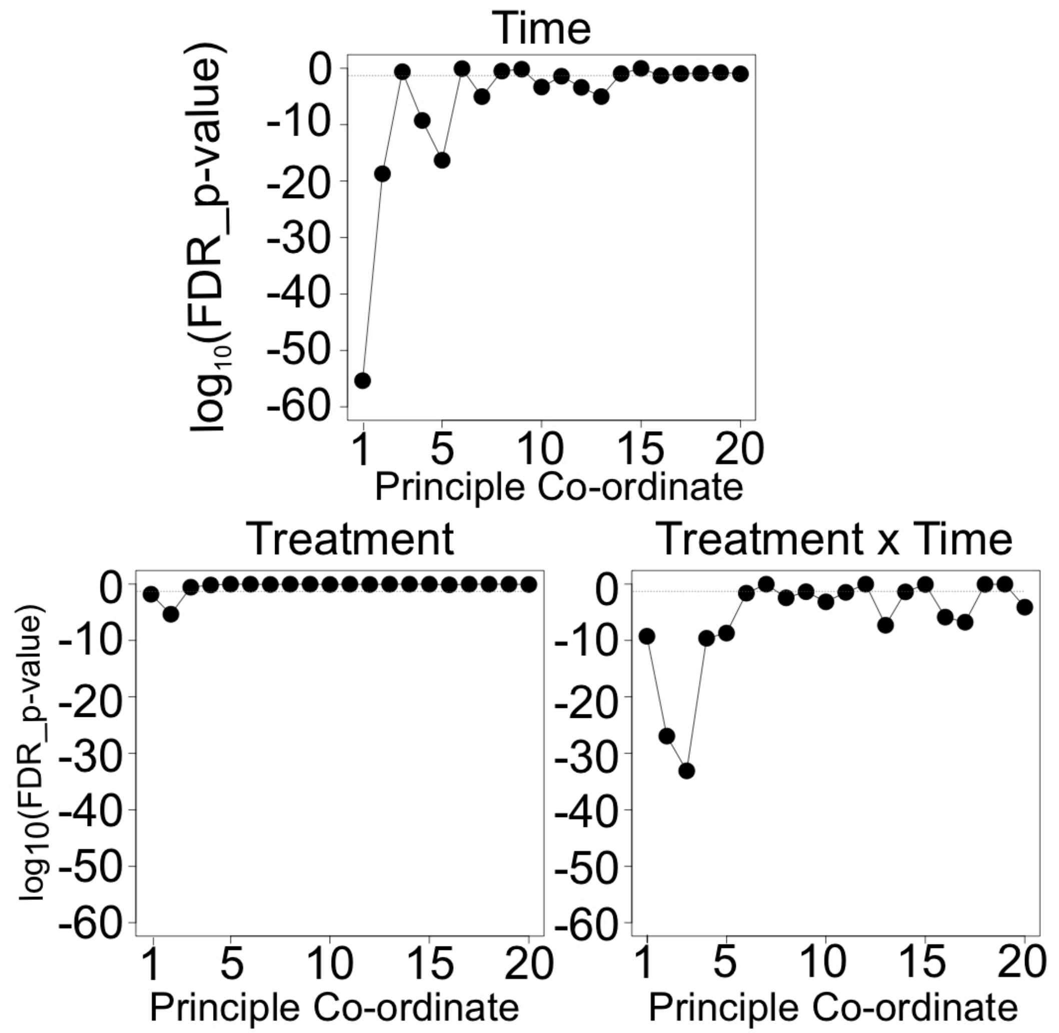


FIGURE 3.S3: Gavage vs Acquired study. Results of a mixed effect linear model with PCoA co-ordinates as the dependent variable. The effects of Time, Treatment (Gavage or Acquired), and the Treatment x Time interaction show significance through the first few PCoA co-ordinates.
Dotted line represents P = 0.05.

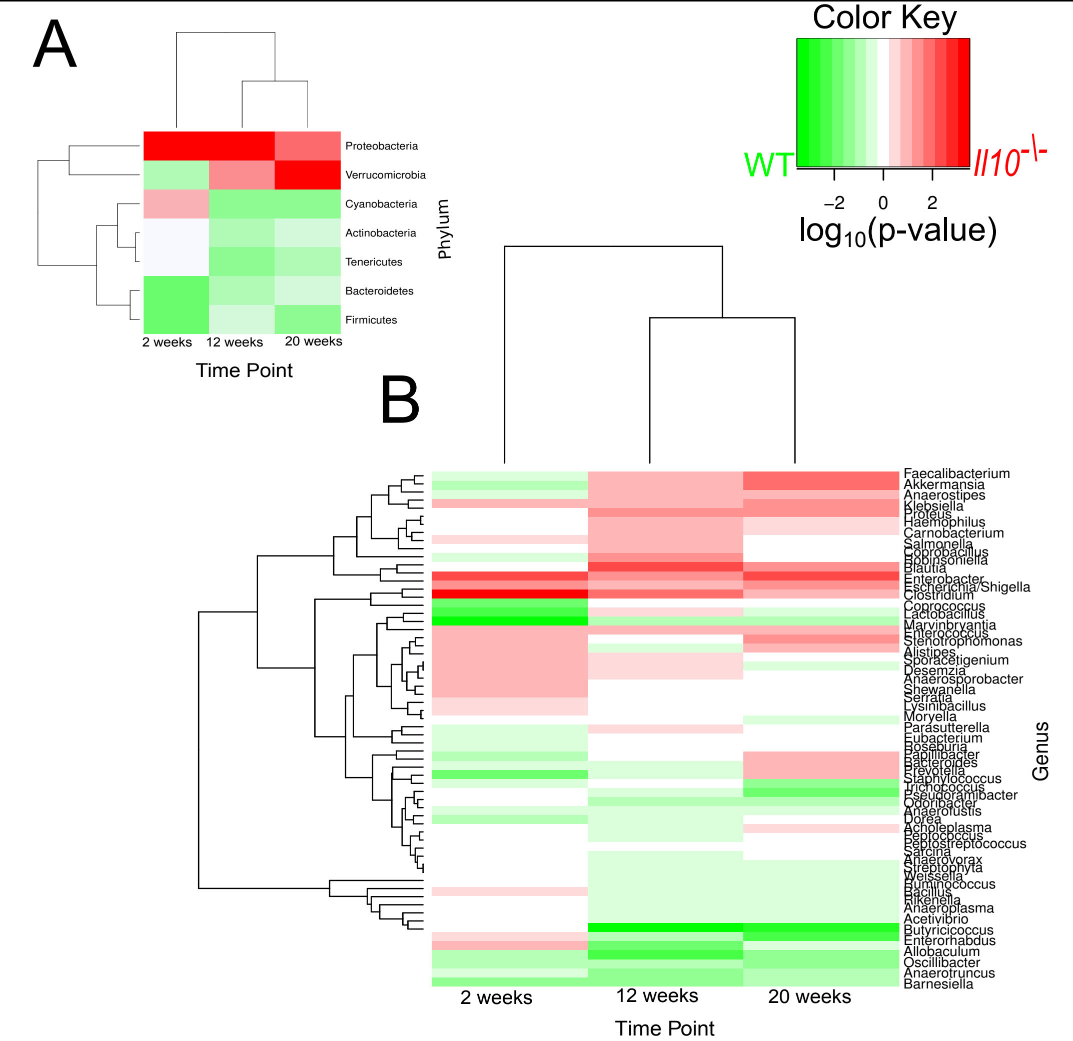


FIGURE 3.S4: Significance assigned through a t-test to differences between WT and *Il10*$^{-\backslash-}$ mice at the phylum (A) and the genus level (B). As seen in the abundance pie chart (Fig 1) an expansion in Proteobacteria occurs early on (A) and while decreasing over time is still significantly different at 20 weeks.

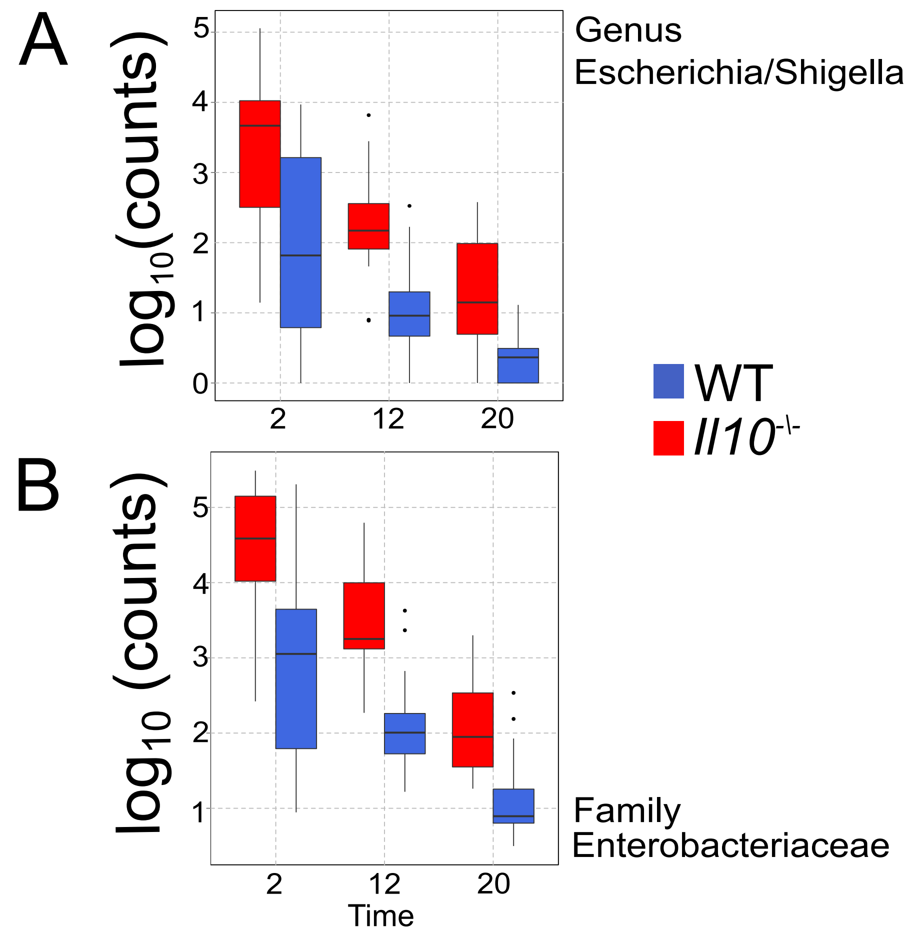


FIGURE 3.S5: WT vs *Il10*$^{-\backslash-}$ study. Log normalized counts of Genus Escherichia/Shigella (A) and Family Enterobactriaceae (B) illustrates the effect time has on the abundance of potential pernicious microbes between healthy (WT) and disease states (*Il10*$^{-\backslash-}$).

Supplemental Tables

TABLES 3.S1: WT vs *Il10*$^{-\backslash-}$ study. Results of the mixed effect linear model conducted in SAS on the first 20 principle co-ordinates of a Bray-Curtis PCoA.

effect	var	NumDF	DenDF	FValue	raw_p	fdr_p
time	axis2	2	103	92.17	<.0001	6.31E-22
time	axis4	2	103	90.45	<.0001	6.31E-22
genotype*time	axis4	2	103	65.76	<.0001	6.15E-18
time	axis1	2	103	65.68	<.0001	6.15E-18
genotype*time	axis3	2	103	37.82	<.0001	5.81E-12
time	axis3	2	103	24.22	<.0001	2.40E-08
genotype*time	axis6	2	103	17.66	<.0001	2.17E-06
time	axis7	2	103	12.7	<.0001	8.78782E-05
genotype	axis1	1	9	43.37	0.0001	0.000672729
genotype	axis2	1	9	20.82	0.0014	0.008164142
time	axis9	2	103	5.92	0.0037	0.019939539
genotype*time	axis11	2	103	5.83	0.004	0.019939539
genotype*time	axis2	2	103	4.22	0.0174	0.080178592
time	axis8	2	103	3.83	0.0248	0.106414406
genotype*time	axis8	2	103	3.35	0.039	0.155964185
genotype*time	axis12	2	103	3.11	0.0487	0.182673722
genotype	axis18	1	9	4.75	0.0572	0.194594529
time	axis15	2	103	2.92	0.0584	0.194594529
time	axis17	2	103	2.86	0.0621	0.196070927
time	axis10	2	103	2.71	0.0716	0.21469964
genotype*time	axis17	2	103	2.6	0.079	0.225790574
genotype*time	axis9	2	103	2.41	0.095	0.259114762
genotype*time	axis15	2	103	2.19	0.1174	0.306201466
genotype	axis9	1	9	2.64	0.1387	0.338170326
genotype*time	axis14	2	103	2	0.1409	0.338170326
genotype	axis6	1	9	2.21	0.1709	0.394361813
genotype*time	axis16	2	103	1.75	0.1794	0.398662011
genotype*time	axis5	2	103	1.58	0.2111	0.452396609
genotype*time	axis19	2	103	1.43	0.2448	0.506556459
genotype	axis10	1	9	1.45	0.2587	0.517458994
genotype	axis3	1	9	1.2	0.3008	0.582250191
genotype*time	axis7	2	103	0.99	0.3758	0.692706328
genotype	axis7	1	9	0.8	0.3957	0.692706328
time	axis6	2	103	0.9	0.4113	0.692706328
genotype	axis15	1	9	0.74	0.4124	0.692706328
genotype*time	axis13	2	103	0.87	0.4212	0.692706328
genotype	axis5	1	9	0.69	0.4272	0.692706328
genotype	axis16	1	9	0.53	0.487	0.759110724
time	axis12	2	103	0.71	0.4934	0.759110724
genotype*time	axis1	2	103	0.67	0.5121	0.768123201
genotype	axis8	1	9	0.4	0.5436	0.795563263
genotype	axis20	1	9	0.35	0.5704	0.814917856

TABLE 3.S1 (continued)						
time	axis13	2	103	0.51	0.6032	0.84162906
time	axis20	2	103	0.47	0.6242	0.843107322
time	axis5	2	103	0.45	0.6377	0.843107322
time	axis14	2	103	0.44	0.6464	0.843107322
genotype*time	axis10	2	103	0.4	0.6743	0.860773517
genotype	axis19	1	9	0.17	0.6905	0.863115976
time	axis18	2	103	0.25	0.7787	0.926819876
genotype	axis13	1	9	0.08	0.7876	0.926819876
genotype*time	axis20	2	103	0.24	0.7878	0.926819876
genotype	axis11	1	9	0.05	0.836	0.964593392
time	axis19	2	103	0.15	0.8566	0.967744314
genotype	axis17	1	9	0.02	0.883	0.967744314
time	axis16	2	103	0.12	0.8888	0.967744314
time	axis11	2	103	0.1	0.9032	0.967744314
genotype*time	axis18	2	103	0.06	0.939	0.978710199
genotype	axis4	1	9	0	0.9481	0.978710199
genotype	axis14	1	9	0	0.9624	0.978710199
genotype	axis12	1	9	0	0.9887	0.988655299

TABLE 3.S2: WT vs *Il10*$^{-\backslash-}$ study. Results of the mixed effect linear model conducted in SAS using Phylum classification with an 80% RDP threshold.

effect	var	Num DF	Den DF	FValue	raw_p	fdr_p
time	Verrucomicrobia	2	103	212.53	<.0001	5.82E-36
time	Bacteroidetes	2	103	110.33	<.0001	2.58E-25
time	Actinobacteria	2	103	51.27	<.0001	2.46E-15
time	Proteobacteria	2	103	45.46	<.0001	3.71E-14
time	Cyanobacteria	2	103	18.73	<.0001	4.84E-07
time	Firmicutes	2	103	12.12	<.0001	6.58417E-05
genotype*time	Verrucomicrobia	2	103	10.53	<.0001	0.000207285
genotype	Proteobacteria	1	9	32.41	0.0003	0.00077909
time	Tenericutes	2	103	7.31	0.0011	0.0025076
genotype	Verrucomicrobia	1	9	15.58	0.0034	0.007077512
genotype*time	Proteobacteria	2	103	3.58	0.0314	0.059944219
genotype*time	Bacteroidetes	2	103	2.86	0.0619	0.108307287
genotype	Bacteroidetes	1	9	1.87	0.2046	0.326642322
genotype*time	Cyanobacteria	2	103	1.55	0.2178	0.326642322
genotype*time	Firmicutes	2	103	1.34	0.2653	0.371432207
genotype	Cyanobacteria	1	9	0.62	0.4525	0.593921596
genotype	Firmicutes	1	9	0.42	0.5335	0.659016473
genotype*time	Tenericutes	2	103	0.46	0.631	0.736182878
genotype	Tenericutes	1	9	0.16	0.6983	0.771840782
genotype	Actinobacteria	1	9	0.11	0.7458	0.783043316
genotype*time	Actinobacteria	2	103	0.17	0.8473	0.847275483

TABLES 3.S3: WT vs *Il10*$^{-\backslash -}$ study. Results of the mixed effect linear model conducted in SAS using Genus classification with an 80% RDP threshold.

effect	**var**	**Num DF**	**Den DF**	**FValue**	**raw_p**	**fdr_p**
time	Barnesiella	2	103	240.5247432	1.54511E-39	2.54944E-37
time	Lactobacillus	2	103	229.1623352	1.19286E-38	9.84109E-37
time	Trichococcus	2	103	218.963221	8.026E-38	4.4143E-36
time	Akkermansia	2	103	163.5932156	1.0661E-32	4.39765E-31
time	Stenotrophomonas	2	103	126.3224669	1.92203E-28	6.34271E-27
time	Bacillus	2	103	88.16302783	4.8571E-23	1.3357E-21
time	Enterococcus	2	103	86.40191919	9.33687E-23	1.9749E-21
time	Enterobacter	2	103	86.33442363	9.57527E-23	1.9749E-21
time	Enterorhabdus	2	103	72.54492974	2.18149E-20	3.9994E-19
time	Bacteroides	2	103	67.44931831	1.89205E-19	3.12187E-18
time	Prevotella	2	103	53.78974686	1.01202E-16	1.51803E-15
time	Klebsiella	2	103	52.69603924	1.73275E-16	2.38253E-15
time	Peptostreptococcus	2	103	48.84490861	1.20521E-15	1.52969E-14
time	Clostridium	2	103	42.20302892	4.09978E-14	4.73162E-13
time	Escherichia_Shigella	2	103	42.1156914	4.30147E-14	4.73162E-13
time	Pseudoramibacter	2	103	38.52889505	3.21665E-13	3.31717E-12
time	Alistipes	2	103	33.60268426	5.83438E-12	5.66278E-11
time	Weissella	2	103	22.03280389	1.08217E-08	9.91986E-08
time	Blautia	2	103	21.76773123	1.30336E-08	1.13187E-07
time	Streptophyta	2	103	20.17125446	4.05326E-08	3.34394E-07
time	Anaerosporobacter	2	103	19.26109297	7.82806E-08	6.15061E-07
time	Anaerostipes	2	103	19.02399081	9.30514E-08	6.97886E-07
time	Lysinibacillus	2	103	15.00373545	1.91206E-06	1.37169E-05
time	Proteus	2	103	14.78227415	2.27042E-06	1.56091E-05
time	Robinsoniella	2	103	14.63100644	2.55392E-06	1.642E-05
genotype*time	Faecalibacterium	2	103	14.61428714	2.5874E-06	1.642E-05
genotype*time	Prevotella	2	103	14.33883245	3.20806E-06	1.96048E-05
time	Staphylococcus	2	103	11.57699778	2.91551E-05	0.000171807
genotype*time	Akkermansia	2	103	11.1114828	4.26965E-05	0.000242928
genotype*time	Bacteroides	2	103	10.78556061	5.58628E-05	0.000307245
genotype*time	Proteus	2	103	9.701522824	0.00013798	0.000734412
time	Marvinbryantia	2	103	9.327027254	0.000189277	0.000975959
genotype*time	Blautia	2	103	9.161419912	0.000217809	0.001089043
time	Ruminococcus	2	103	8.712598421	0.000319281	0.001549451
genotype*time	Dorea	2	103	8.656184634	0.000335072	0.001579624
genotype*time	Alistipes	2	103	8.184041831	0.000502775	0.002304385
time	Acetivibrio	2	103	8.027535536	0.000575573	0.002480913
time	Anaerovorax	2	103	8.014212752	0.000582246	0.002480913
genotype*time	Marvinbryantia	2	103	8.006002933	0.000586398	0.002480913
genotype*time	Anaerostipes	2	103	7.862397382	0.0006641	0.002739413
genotype	Allobaculum	1	9	25.57003433	0.000684112	0.002753135
time	Desemzia	2	103	7.677690605	0.000779713	0.003063157

TABLE 3.S3 (continued)

genotype*time	Coprococcus	2	103	7.479173822	0.000927022	0.003557177
genotype	Barnesiella	1	9	21.83923583	0.001163243	0.004326318
genotype	Clostridium	1	9	21.68210498	0.001191268	0.004326318
genotype*time	Robinsoniella	2	103	7.178529696	0.001206125	0.004326318
time	Anaerotruncus	2	103	7.050383404	0.001349866	0.004700688
time	Odoribacter	2	103	7.035651741	0.001367473	0.004700688
genotype	Marvinbryantia	1	9	20.2240091	0.001495146	0.005034676
genotype*time	Anaerofustis	2	103	6.774744053	0.001721207	0.005679982
time	Haemophilus	2	103	6.603328134	0.002003182	0.006480882
time	Anaerofustis	2	103	6.441697762	0.002312196	0.007336776
genotype	Enterobacter	1	9	16.56747863	0.002798317	0.00871174
genotype*time	Barnesiella	2	103	6.135917426	0.003036491	0.009278166
time	Faecalibacterium	2	103	6.101798397	0.003130517	0.009391552
genotype*time	Sarcina	2	103	5.36804334	0.006058367	0.017850546
genotype*time	Clostridium	2	103	4.925144077	0.009062112	0.025883027
time	Acholeplasma	2	103	4.920780656	0.009098276	0.025883027
genotype	Escherichia_Shigella	1	9	10.85577378	0.009306639	0.02602704
genotype*time	Odoribacter	2	103	4.589945282	0.012316688	0.033870893
genotype*time	Enterorhabdus	2	103	4.535560608	0.012947641	0.035022308
genotype	Acetivibrio	1	9	9.364508083	0.013568366	0.036109361
genotype	Akkermansia	1	9	9.190777485	0.014209147	0.037214433
genotype*time	Lactobacillus	2	103	4.223897485	0.017255829	0.044487685
genotype	Oscillibacter	1	9	8.338075792	0.017955123	0.04557839
genotype*time	Haemophilus	2	103	4.131176608	0.018801074	0.047002685
genotype*time	Coprobacillus	2	103	4.043555319	0.020390976	0.050216583
genotype	Butyricicoccus	1	9	7.500159714	0.022898402	0.055562299
genotype*time	Acetivibrio	2	103	3.901889902	0.023257147	0.055614916
time	Carnobacterium	2	103	3.84800967	0.024452253	0.057637453
genotype	Carnobacterium	1	9	7.115499795	0.025728081	0.059790612
genotype*time	Roseburia	2	103	3.661982998	0.029081646	0.066235475
genotype	Anaerotruncus	1	9	6.698313271	0.02930418	0.066235475
genotype	Proteus	1	9	6.487889854	0.031341793	0.069883727
time	Dorea	2	103	3.53838023	0.032643117	0.071814857
time	Allobaculum	2	103	3.519876189	0.033213332	0.072107891
genotype	Dorea	1	9	6.166432374	0.03480648	0.073965895
time	Salmonella	2	103	3.464973356	0.034965696	0.073965895
genotype	Blautia	1	9	5.946666018	0.037451849	0.078222215
genotype*time	Rikenella	2	103	3.125693539	0.048098181	0.099202499
genotype*time	Pseudoramibacter	2	103	3.040883147	0.052105162	0.106140145
genotype*time	Lysinibacillus	2	103	3.002695001	0.05401899	0.108631538
time	Roseburia	2	103	2.97578501	0.055410515	0.108631538
genotype	Haemophilus	1	9	4.821372809	0.055724288	0.108631538
genotype	Enterococcus	1	9	4.780037149	0.056587779	0.108631538
time	Shewanella	2	103	2.948100986	0.056880228	0.108631538
time	Sporacetigenium	2	103	2.940725488	0.057278447	0.108631538
genotype	Weissella	1	9	4.404265457	0.065262768	0.12236769

TABLE 3.S3 (continued)						
time	Coprobacillus	2	103	2.760106497	0.067970309	0.12601237
genotype	Serratia	1	9	3.816038193	0.082505822	0.151260673
time	Papillibacter	2	103	2.542532553	0.083595734	0.151574682
genotype	Odoribacter	1	9	3.643943093	0.088615834	0.158930572
genotype	Sporacetigenium	1	9	3.458150762	0.095872872	0.17009703
genotype	Pseudoramibacter	1	9	3.208826058	0.106847863	0.1875521
time	Sarcina	2	103	2.267114983	0.108757822	0.188895164
genotype*time	Staphylococcus	2	103	2.226018126	0.113126044	0.192598711
genotype*time	Trichococcus	2	103	2.225108777	0.113224697	0.192598711
genotype	Staphylococcus	1	9	3.013660415	0.116583447	0.196288456
time	Serratia	2	103	2.167729178	0.119630382	0.19938397
time	Coprococcus	2	103	2.130666849	0.123963162	0.204539218
genotype	Klebsiella	1	9	2.842219691	0.126096586	0.205999373
genotype*time	Carnobacterium	2	103	1.992313163	0.141600995	0.229060433
genotype	Shewanella	1	9	2.506037038	0.147870581	0.236880056
genotype*time	Bacillus	2	103	1.908177832	0.153558054	0.243625759
time	Peptococcus	2	103	1.843405789	0.163461005	0.256867293
genotype*time	Klebsiella	2	103	1.797098287	0.170938033	0.266082788
time	Rikenella	2	103	1.727266512	0.182878594	0.282009046
genotype	Anaerofustis	1	9	1.935681454	0.197567292	0.301838918
genotype	Streptophyta	1	9	1.790751056	0.213658616	0.32342818
genotype	Bacillus	1	9	1.736594982	0.22012646	0.33018969
genotype	Faecalibacterium	1	9	1.701738999	0.224430871	0.333613457
genotype	Enterorhabdus	1	9	1.618321828	0.235210198	0.346515024
genotype	Peptostreptococcus	1	9	1.579174229	0.240514758	0.351194116
time	Butyricicoccus	2	103	1.430593211	0.243878822	0.352982506
genotype	Anaerovorax	1	9	1.519534021	0.248918544	0.357143997
time	Moryella	2	103	1.327854783	0.269543625	0.38340257
time	Eubacterium	2	103	1.299706656	0.277044509	0.390703794
genotype*time	Anaerosporobacter	2	103	1.215472346	0.300787669	0.420592927
genotype	Coprococcus	1	9	1.157231668	0.310031306	0.42987534
genotype*time	Stenotrophomonas	2	103	1.158337363	0.318063954	0.437337937
genotype	Stenotrophomonas	1	9	1.090193679	0.323660981	0.441355884
genotype*time	Peptococcus	2	103	1.128309342	0.327545947	0.442992469
genotype	Roseburia	1	9	1.007616535	0.341697834	0.458375143
genotype	Ruminococcus	1	9	0.954696953	0.354057502	0.471124902
genotype	Bacteroides	1	9	0.897679937	0.368148019	0.485955386
genotype*time	Serratia	2	103	0.981956985	0.378056071	0.495073426
genotype*time	Acholeplasma	2	103	0.94769202	0.390988191	0.500194495
genotype	Lysinibacillus	1	9	0.808553985	0.391970847	0.500194495
genotype	Alistipes	1	9	0.805271914	0.392893751	0.500194495
genotype	Rikenella	1	9	0.801026518	0.394092632	0.500194495
genotype*time	Sporacetigenium	2	103	0.916630785	0.403100686	0.507722238
genotype*time	Enterobacter	2	103	0.893164815	0.412504513	0.515630641
genotype*time	Anaerotruncus	2	103	0.870695557	0.421718529	0.523184642
time	Oscillibacter	2	103	0.833124264	0.437596829	0.535287871

TABLE 3.S3 (continued)

genotype*time	Butyricicoccus	2	103	0.832274769	0.437962804	0.535287871
genotype	Anaerosporobacter	1	9	0.580432219	0.465638156	0.564928645
genotype	Lactobacillus	1	9	0.554915671	0.475317949	0.572463223
genotype*time	Peptostreptococcus	2	103	0.719664	0.489347678	0.585089615
genotype*time	Enterococcus	2	103	0.703424562	0.497249225	0.590259871
genotype*time	Ruminococcus	2	103	0.658098061	0.519998811	0.612566736
genotype*time	Anaerovorax	2	103	0.651367786	0.52346612	0.612566736
genotype*time	Salmonella	2	103	0.642308932	0.528170314	0.613719027
genotype*time	Desemzia	2	103	0.601096006	0.550122038	0.634756197
genotype	Desemzia	1	9	0.339004565	0.574698196	0.655487691
genotype	Moryella	1	9	0.336597878	0.576034638	0.655487691
genotype*time	Shewanella	2	103	0.517602351	0.597489505	0.675244989
genotype	Prevotella	1	9	0.281592542	0.608502225	0.683012702
genotype	Peptococcus	1	9	0.268038308	0.617135034	0.688022167
genotype*time	Allobaculum	2	103	0.4638749	0.630149469	0.697816526
genotype*time	Papillibacter	2	103	0.419998065	0.658168228	0.723985051
genotype*time	Escherichia_Shigella	2	103	0.357235352	0.700471183	0.765415531
genotype	Trichococcus	1	9	0.140535462	0.716430645	0.777704319
genotype	Papillibacter	1	9	0.120737618	0.736223495	0.793966514
genotype	Salmonella	1	9	0.107424713	0.750591167	0.804204822
genotype*time	Eubacterium	2	103	0.263162336	0.769132329	0.81875377
genotype	Sarcina	1	9	0.082187197	0.780845282	0.825894049
genotype*time	Weissella	2	103	0.210213188	0.810758275	0.852070799
genotype	Eubacterium	1	9	0.051840995	0.824978566	0.861528249
genotype*time	Streptophyta	2	103	0.172453	0.841840338	0.868675666
genotype*time	Moryella	2	103	0.171843172	0.842352161	0.868675666
genotype*time	Oscillibacter	2	103	0.140091372	0.869444153	0.891045249
genotype	Anaerostipes	1	9	0.016689286	0.90005068	0.916718286
genotype	Robinsoniella	1	9	0.009607163	0.924067705	0.932613463
genotype	Coprobacillus	1	9	0.008886546	0.926961261	0.932613463
genotype	Acholeplasma	1	9	0.00266583	0.959949982	0.959949982

TABLE 3.S4: WT vs *Il10*$^{-\backslash-}$ study. Results of the mixed effect linear model conducted in SAS for Richness rarified to 19,226 sequences. *P < 0.05.

effect	**var**	**NumDF**	**DenDF**	**FValue**	**raw_p**	**fdr_p**
time	richness19226	2	103	7.7	0.0008	0.001648427*
genotype	richness19226	1	9	22.22	0.0011	0.001648427*
genotype*time	richness19226	2	103	1.19	0.3093	0.309285384

TABLE 3.S5: Gavage vs Acquired study. Results of the mixed effect linear model conducted in SAS on the first 20 principle co-ordinates of a Bray-Curtis PCoA.

effect	var	NumDF	DenDF	FValue	raw_p	fdr_p
time	axis1	3	82	222.67	<.0001	1.44E-37
treatment*time	axis2	3	82	75.89	<.0001	3.70E-22
time	axis2	3	82	46.61	<.0001	2.00E-16
treatment*time	axis4	3	82	22.52	<.0001	1.34E-09
time	axis4	3	82	20.24	<.0001	7.08E-09
time	axis7	3	82	14.14	<.0001	1.48E-06
treatment*time	axis13	3	82	11.55	<.0001	1.67559E-05
treatment	axis2	1	6	242.45	<.0001	2.99822E-05
treatment*time	axis16	3	82	9.59	<.0001	0.000101378
treatment*time	axis1	3	82	8.69	<.0001	0.000243805
treatment*time	axis17	3	82	7.43	0.0002	0.000906346
treatment*time	axis10	3	82	7.28	0.0002	0.000975655
time	axis13	3	82	6.29	0.0007	0.002834955
treatment*time	axis8	3	82	5.77	0.0013	0.004826499
treatment*time	axis6	3	82	5.59	0.0015	0.005463469
treatment*time	axis20	3	82	5.55	0.0016	0.005463469
time	axis10	3	82	3.95	0.011	0.035016758
treatment	axis1	1	6	12.27	0.0128	0.038341258
time	axis20	3	82	3.3	0.0245	0.069490814
treatment*time	axis14	3	82	3.14	0.0298	0.080447244
time	axis14	3	82	2.88	0.0407	0.10466861
time	axis16	3	82	2.64	0.0547	0.131291307
treatment*time	axis11	3	82	2.62	0.0561	0.131291307
time	axis11	3	82	2.59	0.0584	0.131291307
time	axis12	3	82	2.43	0.0713	0.154051024
time	axis17	3	82	1.79	0.1563	0.322949277
time	axis8	3	82	1.76	0.1615	0.322949277
time	axis18	3	82	1.37	0.2564	0.494439902
time	axis9	3	82	1.27	0.2914	0.537549714
treatment*time	axis9	3	82	1.25	0.2986	0.537549714
time	axis19	3	82	1.09	0.3564	0.620876507
treatment	axis16	1	6	0.69	0.4374	0.738046976
treatment	axis10	1	6	0.51	0.5014	0.820554016
treatment	axis4	1	6	0.43	0.5383	0.830977213
treatment*time	axis18	3	82	0.73	0.5386	0.830977213
treatment	axis12	1	6	0.32	0.5899	0.884810332
treatment	axis7	1	6	0.28	0.6145	0.896907649
treatment	axis20	1	6	0.22	0.655	0.929592464
treatment*time	axis12	3	82	0.47	0.7013	0.929592464
treatment*time	axis19	3	82	0.47	0.7064	0.929592464
treatment	axis13	1	6	0.15	0.7149	0.929592464
time	axis6	3	82	0.41	0.7446	0.929592464

TABLE 3.S5 (continued)						
treatment	axis8	1	6	0.08	0.7811	0.929592464
treatment	axis19	1	6	0.08	0.7933	0.929592464
treatment*time	axis15	3	82	0.3	0.825	0.929592464
treatment	axis18	1	6	0.04	0.8531	0.929592464
treatment	axis6	1	6	0.03	0.8611	0.929592464
treatment	axis17	1	6	0.03	0.8731	0.929592464
treatment	axis14	1	6	0.02	0.9025	0.929592464
treatment	axis15	1	6	0.01	0.9071	0.929592464
treatment*time	axis7	3	82	0.18	0.9099	0.929592464
time	axis15	3	82	0.17	0.9139	0.929592464
treatment	axis11	1	6	0.01	0.9157	0.929592464
treatment	axis9	1	6	0.01	0.9296	0.929592464

TABLE 3.S6: Gavage vs Acquired study. Results of the mixed effect linear model conducted in SAS using Phylum classification with an 80% RDP threshold.

effect	var	NumDF	DenDF	FValue	raw_p	fdr_p
time	Proteobacteria	3	82	63.63	<.0001	4.30E-20
time	Bacteroidetes	3	82	39.07	<.0001	7.95E-15
time	Actinobacteria	3	82	21.71	<.0001	8.80E-10
treatment*time	Bacteroidetes	3	82	21.69	<.0001	8.80E-10
time	Verrucomicrobia	3	82	11.78	<.0001	6.15E-06
treatment*time	Tenericutes	3	82	9.42	<.0001	6.12957E-05
time	Tenericutes	3	82	8.87	<.0001	9.48165E-05
treatment*time	Actinobacteria	3	82	4.31	0.0071	0.016065531
treatment*time	Proteobacteria	3	82	4.12	0.009	0.017957009
treatment*time	Verrucomicrobia	3	82	3.58	0.0174	0.031277647
treatment	Bacteroidetes	1	6	6.31	0.0458	0.074999645
treatment	Firmicutes	1	6	5.09	0.065	0.09440549
treatment*time	Firmicutes	3	82	2.46	0.0682	0.09440549
time	Firmicutes	3	82	2.34	0.0792	0.101785458
treatment	Verrucomicrobia	1	6	2.46	0.168	0.201626077
treatment	Actinobacteria	1	6	1.81	0.2274	0.255840238
treatment	Proteobacteria	1	6	0.09	0.7709	0.816290897
treatment	Tenericutes	1	6	0.04	0.8451	0.845066159

TABLE 3.S7: Gavage vs Acquired study. Results of the mixed effect linear model conducted in SAS using Genus classification with an 80% RDP threshold.

effect	var	NumDF	DenDF	FValue	raw_p	fdr_p
time	Enterobacter	3	82	69.3968422	1.9456E-22	3.21024E-20
time	Anaerosporobacter	3	82	54.44195167	1.83615E-19	1.4945E-17
time	Robinsoniella	3	82	53.65927634	2.71727E-19	1.4945E-17
time	Enterococcus	3	82	52.0772511	6.07077E-19	2.50419E-17
treatment*time	Haemophilus	3	82	51.05317111	1.0302E-18	3.39968E-17
time	Barnesiella	3	82	48.44818065	4.08164E-18	1.027E-16
time	Klebsiella	3	82	48.32681285	4.35698E-18	1.027E-16
time	Butyricicoccus	3	82	47.42262519	7.10928E-18	1.46629E-16
time	Haemophilus	3	82	41.8985784	1.61773E-16	2.96584E-15
time	Escherichia_Shigella	3	82	39.82443589	5.57533E-16	8.8303E-15
time	Trichococcus	3	82	39.73471097	5.88686E-16	8.8303E-15
time	Clostridium	3	82	38.56288878	1.20553E-15	1.57727E-14
treatment*time	Anaerosporobacter	3	82	38.51369389	1.2427E-15	1.57727E-14
time	Blautia	3	82	37.78117344	1.95834E-15	2.30805E-14
time	Odoribacter	3	82	33.84762315	2.46315E-14	2.70946E-13
treatment*time	Trichococcus	3	82	32.87140073	4.73325E-14	4.88116E-13
time	Rikenella	3	82	30.48110808	2.45051E-13	2.37844E-12
treatment*time	Bacteroides	3	82	29.46054087	5.04676E-13	4.6262E-12
time	Faecalibacterium	3	82	28.13375933	1.31604E-12	1.14287E-11
time	Prevotella	3	82	27.88109285	1.58358E-12	1.30646E-11
time	Lactobacillus	3	82	27.52150899	2.0637E-12	1.62148E-11
treatment*time	Citrobacter	3	82	26.81602171	3.48689E-12	2.61517E-11
time	Citrobacter	3	82	26.69362237	3.82167E-12	2.74164E-11
time	Enterorhabdus	3	82	23.58796382	4.19831E-11	2.88634E-10
time	Anaerovorax	3	82	23.37236879	4.9843E-11	3.28964E-10
time	Parabacteroides	3	82	22.03816294	1.46482E-10	9.29595E-10
time	Acetivibrio	3	82	19.36586148	1.38512E-09	8.4646E-09
time	Alistipes	3	82	17.83101083	5.32687E-09	3.13905E-08
time	Ruminococcus	3	82	15.43451366	4.77469E-08	2.71663E-07
time	Bacteroides	3	82	14.94475871	7.58244E-08	4.17034E-07
treatment*time	Lactobacillus	3	82	14.2832667	1.42762E-07	7.59865E-07
time	Allobaculum	3	82	13.12202921	4.43686E-07	2.28776E-06
time	Syntrophococcus	3	82	13.0173465	4.9217E-07	2.46085E-06
time	Dorea	3	82	12.78067602	6.2279E-07	3.02236E-06
treatment*time	Parasutterella	3	82	12.00076591	1.36522E-06	6.43603E-06
time	Bacillus	3	82	11.90438657	1.50576E-06	6.8657E-06
time	Marvinbryantia	3	82	11.88256727	1.53958E-06	6.8657E-06
treatment*time	Syntrophococcus	3	82	11.10636989	3.41832E-06	1.48427E-05
treatment*time	Odoribacter	3	82	10.37803069	7.32282E-06	3.09812E-05
time	Akkermansia	3	82	9.799607981	1.35377E-05	5.58432E-05
treatment*time	Dorea	3	82	9.350239758	2.19495E-05	8.83335E-05
treatment*time	Acholeplasma	3	82	9.262346272	2.41402E-05	9.48364E-05
treatment*time	Robinsoniella	3	82	9.063728335	2.99517E-05	0.000114931
time	Coprococcus	3	82	8.89157956	3.61392E-05	0.000135522

TABLE 3.S7 (continued)

treatment*time	Rikenella	3	82	8.30898407	6.86259E-05	0.000251628
treatment*time	Coprococcus	3	82	7.918550412	0.000106001	0.000380221
time	Acholeplasma	3	82	7.631944627	0.00014623	0.000513359
treatment*time	Pantoea	3	82	7.077789225	0.000274087	0.000942173
time	Weissella	3	82	6.849648915	0.000355836	0.001198222
treatment	Odoribacter	1	6	51.04179796	0.000379007	0.001232999
treatment	Trichococcus	1	6	50.93821908	0.000381109	0.001232999
time	Papillibacter	3	82	6.744403027	0.000401558	0.001274174
treatment*time	Marvinbryantia	3	82	6.65512351	0.000445023	0.001385448
time	Pantoea	3	82	6.508526065	0.000527073	0.001610499
time	Lactococcus	3	82	6.486077747	0.000540935	0.001622806
treatment*time	Anaerotruncus	3	82	6.135544114	0.000812739	0.002394677
treatment	Haemophilus	1	6	34.24934536	0.001098857	0.003180901
treatment	Lactobacillus	1	6	33.89645949	0.001128911	0.003211557
treatment*time	Parabacteroides	3	82	5.748559947	0.001278662	0.003575918
treatment*time	Prevotella	3	82	5.596108501	0.001530188	0.004208016
time	Oscillibacter	3	82	5.537900305	0.00163904	0.004433468
time	Eubacterium	3	82	5.474112909	0.001767411	0.004703593
treatment	Rikenella	1	6	23.40436995	0.002887223	0.007561776
treatment*time	Helicobacter	3	82	4.870195275	0.003626921	0.009350655
treatment*time	Klebsiella	3	82	4.839780721	0.003761503	0.009548432
treatment*time	Blautia	3	82	4.545715404	0.005355808	0.013389519
treatment*time	Acetivibrio	3	82	4.222934524	0.0079108	0.019481821
treatment	Citrobacter	1	6	13.61496127	0.010210601	0.024472583
time	Adlercreutzia	3	82	4.01081051	0.010233989	0.024472583
treatment	Parabacteroides	1	6	13.2823767	0.010776829	0.025402525
time	Carnobacterium	3	82	3.930115039	0.011289699	0.026236625
treatment	Bacteroides	1	6	12.75904312	0.011756095	0.02694105
treatment	Syntrophococcus	1	6	12.53162051	0.012218858	0.027617966
time	Sporacetigenium	3	82	3.833424853	0.012701003	0.028319805
treatment	Prevotella	1	6	11.80588279	0.013869291	0.03051244
treatment*time	Allobaculum	3	82	3.736005153	0.014303765	0.031054226
time	Coprobacillus	3	82	3.696959307	0.015002226	0.032147627
time	Parasutterella	3	82	3.679293126	0.015329474	0.032427733
treatment*time	Coprobacillus	3	82	3.615713721	0.016568086	0.03460423
treatment	Lawsonia	1	6	10.68766418	0.017051303	0.0351667
treatment	Anaerosporobacter	1	6	10.62312145	0.017263653	0.0351667
time	Helicobacter	3	82	3.528245738	0.018439088	0.036915814
time	Lawsonia	3	82	3.522475403	0.018569773	0.036915814
treatment	Alistipes	1	6	9.758390257	0.020484471	0.040237354
treatment	Adlercreutzia	1	6	8.381598427	0.027509828	0.053401431
treatment	Desulfovibrio	1	6	8.205679869	0.028630375	0.054930371
treatment*time	Lactonifactor	3	82	3.071207465	0.032300964	0.061260449
time	Anaerotruncus	3	82	3.003561657	0.035101518	0.065815346
treatment	Allobaculum	1	6	6.934796676	0.038882252	0.071136461
treatment*time	Bifidobacterium	3	82	2.914719539	0.039153417	0.071136461

TABLE 3.S7 (continued)

treatment	Bifidobacterium	1	6	6.857834592	0.039652764	0.071136461
treatment	Anaerotruncus	1	6	6.856730577	0.039663966	0.071136461
treatment	Helicobacter	1	6	6.274395422	0.046222471	0.08200761
treatment*time	Anaerostipes	3	82	2.742749904	0.048377472	0.084917903
treatment*time	Anaerovorax	3	82	2.622596773	0.056084319	0.097409607
treatment*time	Escherichia_Shigella	3	82	2.569702478	0.059854824	0.102875479
treatment	Anaerovorax	1	6	5.236518179	0.062084909	0.10560835
time	Bifidobacterium	3	82	2.442833515	0.069960399	0.117790468
treatment*time	Ruminococcus	3	82	2.371100449	0.07640773	0.127346217
treatment	Pantoea	1	6	4.395506969	0.080862903	0.13342379
treatment*time	Bacillus	3	82	2.316399796	0.0817177	0.133499213
treatment	Enterorhabdus	1	6	4.177714294	0.086967953	0.140683453
treatment*time	Adlercreutzia	3	82	2.200346564	0.094224456	0.15094209
time	Weeksella	3	82	2.114590985	0.104664664	0.166054516
treatment*time	Weeksella	3	82	2.054930651	0.112592469	0.176931022
treatment*time	Carnobacterium	3	82	2.039487203	0.114739098	0.178347354
treatment*time	Butyricicoccus	3	82	2.032980271	0.115655557	0.178347354
treatment	Coprococcus	1	6	3.294803122	0.119418482	0.182225447
treatment	Bacillus	1	6	3.27383635	0.120379235	0.182225447
treatment*time	Akkermansia	3	82	1.964370505	0.125767016	0.188650524
treatment	Anaerostipes	1	6	3.008826356	0.133503908	0.198451755
treatment	Blautia	1	6	2.913493488	0.138712616	0.204353407
treatment*time	Enterococcus	3	82	1.832524287	0.147685127	0.215646424
time	Lactonifactor	3	82	1.816799096	0.150535883	0.217430216
treatment*time	Enterobacter	3	82	1.811316846	0.151542272	0.217430216
treatment*time	Desulfovibrio	3	82	1.758605423	0.161558381	0.22980287
treatment	Papillibacter	1	6	2.51690635	0.163727243	0.230897394
treatment*time	Faecalibacterium	3	82	1.703622969	0.172689413	0.241472484
time	Moryella	3	82	1.681562534	0.177360954	0.24592065
treatment*time	Eubacterium	3	82	1.56896291	0.203165676	0.277475512
treatment	Escherichia_Shigella	1	6	2.0363441	0.203482042	0.277475512
treatment	Coprobacillus	1	6	1.87753749	0.219670254	0.294826962
treatment	Acetivibrio	1	6	1.876518061	0.219780099	0.294826962
treatment	Akkermansia	1	6	1.620485613	0.250121428	0.332822868
treatment*time	Barnesiella	3	82	1.334417651	0.268876827	0.354917411
treatment*time	Weissella	3	82	1.31389027	0.275493783	0.360765668
treatment	Weissella	1	6	1.415032397	0.279156557	0.362683716
treatment	Klebsiella	1	6	1.259801679	0.304585854	0.392630203
treatment	Moryella	1	6	1.193022458	0.316625539	0.404986154
treatment	Dorea	1	6	1.179609607	0.319132097	0.405052277
time	Anaerostipes	3	82	1.114035549	0.348255281	0.438642147
treatment	Enterobacter	1	6	1.014541633	0.352696245	0.439591064
treatment*time	Lawsonia	3	82	1.099107096	0.35433704	0.439591064
treatment	Clostridium	1	6	0.885587153	0.383002504	0.471250163
treatment	Barnesiella	1	6	0.875428538	0.385568315	0.471250163
treatment*time	Shewanella	3	82	1.003512189	0.395577926	0.476510183

TABLE 3.S7 (continued)						
time	Shewanella	3	82	1.003357736	0.395647849	0.476510183
treatment	Weeksella	1	6	0.799628909	0.405643292	0.485008283
treatment	Moritella	1	6	0.770246343	0.413898372	0.491318211
treatment	Sporacetigenium	1	6	0.724446502	0.427346225	0.503658051
time	Desulfovibrio	3	82	0.913064962	0.43840535	0.513027537
treatment	Enterococcus	1	6	0.661602789	0.447062497	0.519474028
treatment	Lactococcus	1	6	0.590709276	0.471303916	0.541641483
treatment	Carnobacterium	1	6	0.580277513	0.475070688	0.541641483
treatment*time	Oscillibacter	3	82	0.83508568	0.478428891	0.541641483
treatment*time	Papillibacter	3	82	0.833504632	0.479270645	0.541641483
treatment	Ruminococcus	1	6	0.535675011	0.491816892	0.552039368
treatment*time	Lactococcus	3	82	0.801312652	0.496673419	0.553723744
time	Moritella	3	82	0.761876547	0.518678896	0.574375959
treatment*time	Moryella	3	82	0.598014841	0.618106776	0.679917454
treatment	Robinsoniella	1	6	0.235206028	0.6448879	0.704678831
treatment	Butyricicoccus	1	6	0.211514226	0.661777354	0.718376733
treatment	Marvinbryantia	1	6	0.147688859	0.714005072	0.768642519
treatment*time	Sporacetigenium	3	82	0.450821552	0.717399685	0.768642519
treatment*time	Enterorhabdus	3	82	0.39504908	0.756893658	0.805725507
treatment*time	Alistipes	3	82	0.338878535	0.797257464	0.843253087
treatment*time	Clostridium	3	82	0.308967098	0.818838034	0.860562265
treatment*time	Moritella	3	82	0.290745955	0.831964522	0.86882371
treatment	Oscillibacter	1	6	0.031076622	0.865869167	0.898543476
treatment	Acholeplasma	1	6	0.025788567	0.877688115	0.905115868
treatment	Eubacterium	1	6	0.022736028	0.885087165	0.907076909
treatment	Lactonifactor	1	6	0.01463053	0.907674326	0.924483109
treatment	Shewanella	1	6	0.006518551	0.938276411	0.949789005
treatment	Faecalibacterium	1	6	7.01496E-05	0.993588905	0.999647374
treatment	Parasutterella	1	6	9.81594E-08	0.999760177	0.999760177

TABLE 3.S8: Gavage vs Acquired study. Results of the mixed effect linear model conducted in SAS for Richness rarified to 2,192 sequences. *P < 0.05.

effect	var	NumDF	DenDF	FValue	raw_p	fdr_p
time	richness2192	3	81	19.11	<.0001	5.49E-09*
treatment	richness2192	1	6	0.72	0.4277	0.641483116
treatment*time	richness2192	3	81	0.06	0.9819	0.981853984

TABLE 3.S9: SAS code used to run mixed linear model for each taxon in order. Other ranks were analyzed using the same method.

```
%MACRO ord (
Var1,
Var2,
Var3,
.
.
Var57);
%DO i=1 %TO 57;

  proc mixed data=ord covtest;

          class  treatment time cage sampleID;

          model &&var&i= treatment time treatment*time / residual outp=r1_&i outpm=r2;

                    random cage(treatment) ;

                    repeated time / subject=mouse type=cs;

                    lsmeans        treatment time treatment*time;

                    ods output Tests3  = overall&i;

                   ods output diffs   = comparison&i;

                   ods output LSMeans = means&i;

                    ods output covParms = cage&i;

   run;
.
.
%END;
%MEND ord;
%ord (
Acetivibrio,
Acholeplasma,
Akkermansia,
.
.
Weissella);
```