University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Dissertations & Theses in Food Science and Technology

Food Science and Technology Department

8-10-2012

ANALYSIS OF MICROBIAL DIVERSITY BY AMPLICON PYROSEQUENCING

Ryan Legge
University of Nebraska-Lincoln, rmlegge@gmail.com

Follow this and additional works at: http://digitalcommons.unl.edu/foodscidiss

Part of the Environmental Microbiology and Microbial Ecology Commons, and the Food Microbiology Commons

Legge, Ryan, "ANALYSIS OF MICROBIAL DIVERSITY BY AMPLICON PYROSEQUENCING" (2012). *Dissertations & Theses in Food Science and Technology.* Paper 25.
http://digitalcommons.unl.edu/foodscidiss/25

This Article is brought to you for free and open access by the Food Science and Technology Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Dissertations & Theses in Food Science and Technology by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

ANALYSIS OF MICROBIAL DIVERSITY BY AMPLICON PYROSEQUENCING

by

Ryan Matthew Legge

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Food Science and Technology

Under the Supervision of Professor Andrew Benson

Lincoln, Nebraska

August, 2012

ANALYSIS OF MICROBIAL DIVERSITY BY AMPLICON PYROSEQUENCING

Ryan Legge, Ph.D.

University of Nebraska, 2012

Advisor: Andrew Benson

Microorganisms numerically dominate terrestrial biodiversity, and play important biochemical and geochemical roles in the environments they inhabit. To understand structure and function of complex ecosystems, it is essential to identify primary drivers of microbial diversity and community structure. Historically, the study of microbial ecology was reductive, limited to microbes able to be cultured and enumerated. Microbes meeting this criterion were thought to comprise the dominating members of the environments they were isolated from, however, estimates suggesting up to 99% of the endogenous species are uncultivable with existing methodologies; a concept that reflects experimental failure, rather than a verifiable conclusion. Therefore surveys of microbial community members relying solely on culture-based techniques will severely underestimate the extent of microbial diversity. Analytical methods for DNA sequencing have progressed over the last 30 years allowing for increasingly detailed analysis of microbial communities. Microbes can be recognized and their function can be understood at the DNA/RNA level without cultivation bias through molecular techniques which analyze content based on microbial DNA isolated from environmental samples. Using high-capacity sequencing, environmental samples can be characterized at resolution, ultimately allowing communities to be compared on the basis of their taxonomic or phylogenetic content as well as on functions the

microbes carry out. In this dissertation research, two unique studies were explored. Studies focus on differences in composition of microbial communities as a phenotype in the GI tract of animals in a genetic selection experiment and as a measure of contamination risk in food production with 454-based pyrosequencing of the 16s rRNA amplicon. These two studies give credence to the applications of “-omics” techniques in addressing questions relevant to fundamental and applied biological disciplines. Ultimately, studies like these are creating paradigm shifts in how we view food production and human health as they begin to uncover the entire microbial community to unparalleled levels. Continued advancements in the technology itself and the associated bioinformatics tools will influence a broad cross-section of problems in food production, heath care, and water and land management.

DEDICATION

To my family - Roger, Barb and Laura - for their infallible support.

ACKNOWLEGEMENTS

Thank you to my family for their continued backing; I don't know who's more relieved to have this degree complete.

Acknowledgement goes to the members of my supervising committee: Dr. Stephen Kachman, Dr. Jens Walter, Dr. Robert Hutkins, Dr. Paul Blum, and Dr. Merlyn Nielsen. Thank you for your advice, suggestions, instructions and warnings.

Dr. Merlyn Nielsen and his lab members, especially Ms. Rhonda Griess, were especially valuable for their care with and use of many mice, resources, and results. Both contributed heavily to the experimental design, procedure, and collection of results in the rearing, antimicrobial studies with the heat loss mice and the donor animals for transfer studies into germ-free mice. Dr. Nielsen also contributed the statistical analysis for the rearing and transfer studies and his student Adrienne Bhatnagar contributed to the compiling and statistical analysis of the many antibiotic perturbation studies conducted by Dr. Nielsen and Ms. Griess.

Dr. Daniel Peterson contributed greatly through the production of germ-free mice used in the transfer experiment. Thank you to his lab and its members, especially Ms. Chris Klintworth and Mr. Robert Schmaltz, for overseeing the transfer and monitoring the feed intake of these animals.

Mr. The Nguyen for creating the sequence database and his help through many computing requests and database modifications to accommodate my sequence output. Also, Mr. Nguyen and Dr. Etsuko Moriyama for their patience and help as I struggled through the use of her servers.

Thank you to past and current members of the lab. Dr. Fanguri Ma for his development of the CLASSIFIER+CD-HIT pipeline, for help in computer programing, and also as a companion.

Dr. Jae Kim who helped establish and maintain pyrosequencing to the high standard upheld in the lab and for his shared contribution to the sequencing completed in this work. Mr. Joseph Nietfeldt for his patience with endless questions and his never failing advice over the years. Dr. Chaomei Zhang and Mrs. Min Zhang, their assistance abetted the foundation of my laboratory training through early pathogen research and aseptic technique. Dr. Zhang also directly contributed through the collection and extraction of the core spinach samples.

Courtney Parker at Chiquita Foods allowing and organizing the shipment of spinach samples for both the core spinach group and in the pooled leaves used in the contamination/abuse study.

Dr. James Wells and Dr. James Bono associated with the USDA-MARC herd for their collection of the core bovine fecal isolates and Dr. Patrick Schnable for supplying samples from the NAM collection and Ms. Denise Zinniel for cultivating and collecting these core corn samples.

Thank you to all friends and colleges for hashing out research, providing welcome distractions, and listening to me vent over these years.

Funding was awarded from the USDA National Needs Fellowship and through the Department of Food Science and Technology.

Lastly, the upmost gratitude to my supervising professor, Dr. Andrew Benson, despite countless commitments and time better spent, you were always available for direction and even mannered critique. Without your counseling and support this dissertation would have been impossible.

TABLE OF CONTENTS

Page

Title …… I

Abstract……..... II

Dedication ……... IV

Acknowledgements ……...……. V

Table of Contents ……...… VII

List of Tables ……..... XI

List of Figures ……..... XII

1. INTRODUCTION AND LITERATURE REVIEW …..... 1

1.1. Evolution and development of microbial diversity……..... 1

1.2. Advances in methodology that promoted exploration of environmental DNA…. 4

1.2.1. DNA extraction method……..... 6

1.2.2. Use of polymerase chain reaction (PCR)……... 7

1.2.3. Development of a molecular clock……. 8

1.3. Pre-sequencing microbial methods……..…….. 11

1.3.1. Denaturation and Temperature Gradient Gel Electrophoresis – DGGE/TGGE…… 12

1.3.2. Terminal restriction fragment length polymorphism – T-RFLP…… 14

1.3.3. Quantitative PCR – qPCR……. 15

1.3.4. Stable isotope probing – SIP……. 16

1.3.5. Nucleic acid hybridization arrays…… 17

1.3.6. Fluorescent in situ hybridization – FISH…… 18

1.4. Sequencing…….. 19

1.4.1. Massively parallel sequencing…… 19

1. The Roche 454 FLX platform…… 21

2. The Illumina (Solexa) GAIIx…… 22

1.4.1.1. Amplicon Sequencing…….. 23

1.4.1.2. Shotgun Sequencing…… 26

1.5. Bioinformatics…….. 28

1.5.1. Database dependent bioinformatics……... 28

1.5.2. Database independent bioinformatics……. 32

1.6. Limitations of Molecular Methods ……. 35

1.7. Applications…… 37

1.7.1. Microbial ecology and population biology…… 37

1.7.1.1. Example: Selective breeding for feed intake characteristics in mice resulting in unique gut microbial communities that contribute to feed intake phenotypes……..... 39

1.7.1.2. Example: Use of pyrosequencing to identify new indicators of fecal contamination and temperature abuse in leafy greens. ……...….. 41

2. SELECTIVE BREEDING FOR FEED INTAKE CHARACTERISTICS IN MICE RESULTS IN UNIQUE GUT MICROBIAL COMMUNITIES THAT CONTRIBUTE TO FEED INTAKE PHENOTYPES…… 43

2.1. Abstract…….. 43

2.2. Introduction……... 44

2.3. Materials and Methods…….. 48

2.3.1. Founder experimental animals and selection process…….. 48

2.3.2. Experimental animals, survey protocols, and sample collection…….. 50

2.3.2.1. Line-replicate-generation experiment…..……. 50

2.3.2.2. Antibiotic perturbation animals……..... 50

2.3.2.3. Transplant and gnotobiotic animals…… 51

2.3.2.4. Co-mingled animals…… 52

2.3.3. Pyrosequencing…… 53

2.3.4. Raw data filtering and binning…….. 54

2.3.5. Taxonomic analysis and statistical methods…… 55

2.3.6. Antibiotic perturbation…….. 56

2.3.7. Transplantation of microbiota into germ-free animals…… 57

2.4. Results…… 58

2.4.1. Effects of rearing (Co-mingling) on the feed intake trait…… 58

2.4.2. Differentiation of the gut microbiota between the MH and ML selection lines……. 59

2.4.3. Antimicrobial modification of the line specific microbiota……..... 62

2.4.4. Transfer of gut microbiota from MH and ML lines to germ-free animals…….. 63

2.5. Discussion…… 65

Author Page…… 78

3. USE OF PYROSEQUENCING TO IDENTIFY NEW IDNICATORS OF FECAL CONTAMINATION AND TEMPERATURE ABUSE IN LEAFY GREENS…….. 79

3.1. Abstract……. 79

3.2. Introduction…….. 80

3.3. Materials and Methods…… 84

3.3.1. Experimental samples, survey protocols and sample collection……. 84

3.3.1.1. Spinach samples…… 84

3.3.1.2. Bovine fecal samples……. 85

3.3.1.3. Corn samples…….. 85

3.3.1.4. Sampling for temperature abuse and cross contamination of feces on spinach…… 85

3.3.2. Whole DNA extraction……. 86

3.3.3. Pyrosequencing……..... 87

3.3.4. Raw data filtering and binning……. 88

3.3.5. Taxonomic analysis and statistical method…….. 89

3.4. Results…… 91

3.4.1. Technical Repeats of Spinach Composites……..…….. 91

3.4.2. Core microbial composition differences of spinach and corn leaves to bovine feces…… 92

3.4.3. Contamination and abuse of spinach…… 93

3.4.3.1. Detection of bovine feces on spinach …….….. 94

3.4.3.2. Temperature abuse of raw spinach leaves…….. 96

3.5. Discussion…… 97

Author Page…… 109

4. DISSCUSSION…….. 110

LITERATURE CITED …….. 112

LIST OF TABLES

SELECTIVE BREEDING FOR FEED INTAKE CHARACTERISTICS IN MICE RESULTS IN UNIQUE GUT MICROBIAL COMMUNITIES THAT CONTRIBUTE TO FEED INTAKE PHENOTYPES.

Table 1. Feed intake per body weight of animals reared in like lines or co-mingled across lines……..... 69

Table 2. Descriptive statistics for principle taxa……... 70

Table 3. Taxonomic differences between lines across two different independent repetitions selected for heat loss characteristics.……...…… 73

Table 4. Taxonomic correlations across one replicate for feed intake characteristics….. 74

Table 5. Feed intake per body weight of animals under antibiotic treatment……..…… 75

Table 6. Feed intake per body weight of animals of gnotobiotic animals before and after line specific microbiome colonization.…… 76

LIST OF FIGURES

SELECTIVE BREEDING FOR FEED INTAKE CHARACTERISTICS IN MICE RESULTS IN UNIQUE GUT MICROBIAL COMMUNITIES THAT CONTRIBUTE TO FEED INTAKE PHENOTYPES.

Figure 1. Effects of caging environments on feed intake of animals sleected for heat loss characteristics…….. 69

Figure 2 (A,B). Effects of antimicrobials on feed intake in mice selected for heat los... 75

Figure 3. Transfer of feed intake traits into Germ-Free recipients using fecal material from donor MH and ML selection lines……... 76

Figure 4 (A,B). Box and Whisker plots of *L. apodemi* levels in untreated and antimicrobial treated lines…….……... 77

ASSESSING THE MICROBIAL DIVERSITY ON THE SPINACH EPIPHYTIC SURFACE AND INDETIFYING POTENTIAL MICROBIAL INDICATORS FROM FECAL CONTAMINATION AND TEMPEATURE ABUSE.

Figure 1. Pairwise combinations of data from four spinach biological replicates... …… 102

Figure 2. Principle component analysis grouping of species level taxa between bovine feces, corn leaves, and spinach leaves……..……..... 103

Figure 3. Box and whisker plots of bovine feces, corn leaves, and spinach leaves…….. 104

Figure 4 (1, 2, 3). Species richness of contaminated verse non-contaminated spinach samples……... 105

Figure 5. Principle component analysis grouping of species level taxa between bovine feces, spinach leaves, and pooled fecal contaminated or uncontaminated spinach leaves……. 106

Figure 6. Principle component analysis grouping of species level taxa between spinach leaves, and temperature abused or un-abused spinach leaves……. 107

Figure 7. Effects of contamination/abuse on the spinach epiphyte……. 108

1. INTRODUCTION AND LITERATURE REVIEW

1.1. Evolution and development of microbial diversity.

Evolution of prokaryotes has occurred over the last 3.8 billion years in responses to varying environmental conditions (1). Due to short generation time, metabolic flexibility and ability to acquire genomic information across phylogenetic barriers nearly every terrestrial environment with conditions to sustain life contains microbes. The biodiversity dominated by microorganism can regulate behavior and function of the environments they inhabit. For example, baterioplankton are thought to be the primary consumption method for regulating amount of dissolved organic carbon in the Sargasso Sea (2). Microorganisms can perform functions useful to their host, as bacteria can outnumber host cells by an order of magnitude (3), the metabolic function encoded by bacteria in the human gut is equal to that of a virtual organ (4). To understand form and function of complex ecosystems identifying primary drivers for microbial diversity and community structure is essential. Since Darwin published his natural selection theories, a fundamental idea in ecology was that spices were motivated by competitive exclusion, later this theory was applied to microbial eclogy. This theory implies that two species competing for the same resource will be unable to coexist and the species more capable of acquiring limited resources will drive out inferior competitors needing that resource (5, 6). Differences in the competitive ability between species cause the abundance of competitors to diverge over time, with the better competing species becoming more common (6, 7). However, while routinely shown in laboratory settings (6) removal of all but the one highly fit

competitor is rarely seen in natural environments. Species coexist by occupying different niches (8, 9). Niches define how species interact with complex environments, species differences and their specific, unique interactions with the environment maintain diversity and prevent competitive exclusion (7). The strength of niche differences versus competition determines if species will coexist. If the niche differences are strongest, species will coexist. However, if competitive differences are strongest exclusion will occur across insignificant niche differences (10). How strongly niche differences stabilize coexistence in natural environments has been an intriguing question and theories have begun to address how microbial assemblages are maintained.

The insurance hypothesis (11) is one attempt to explain microbial diversity in ecosystems. The hypothesis suggests that high levels of diversity protect communities from catastrophic events in the environment that might otherwise cause the system to fail. A diverse subpopulation of organisms in this instance allows for long term success as it broadens the sustainable conditions in which community can endure (12).

Diversity is therefore desirable for ecosystem stability as it provides functional redundancy as a means to protect key processes for the community survival (13). However diversity itself is not the driver of system assembly and function as the system itself must have the ability to comprise and sustain certain species or functional properties to lead to stability in the ecosystem (14).

In other environments, such as the mammalian gut, populations are remarkably stable within individuals (15). Environments within these host “super-organisms”

attempt promote a core group of microbial populations while limiting blooms of subpopulations detrimental to the host (13). Within vertebrate hosts having a microbial core seems to be a universal feature (16, 17) and the establishments of correct assemblage of microbial communities impacts host fitness (18-20). Still functional redundancy is an important driver even in a system with low diversity. No matter the extent of microbial diversity the collective microbial communities need to provide enough genetic redundancy and transcriptome diversity to offer resilience to the gut ecosystem (13). If individual species can contribute a wide range of required responses, the community as a whole needs less diversity to maintain stability. (11)

In contrast, neutral theory assumes all species ecologically identical and niche differences fail to explain diversity. The theory assumes that the differences between community members of tropically similarity are irrelevant to their success in the environment or "neutral". Highly diverse communities are maintained because chance extinctions are balanced by new speciation, and random changes in community structure over time are due to ecological drift or from outside interaction being introduced in the community. But changes in the community structure arise in an unpredictable way (21). In this theory once a community is at capacity a new individual can only establish if adjacent individual becomes extinct or the makes a new space for the new community member. (21, 22)

Neutral theory predicts that all species in a community have one shared niche, where other theories believe that species have unique no-overlapping niches. In nature neither of these extremes is accurate but rather a balance between these events is thought to closely relate to the natural environments' balance. (22)

1.2. Advances in methodology that promoted exploration of environmental DNA.

Though microbially-dominated ecosystems have been known for over a century, lack of proper methods for identifying and enumerating microorganism has limited our understanding of their behavior in natural ecosystem (23). There is a need for more accurate assessment of natural microbial ecosystems. This can be accomplished through culture independence, where details may come from following key species and coherent phylogenetic groups of microorganisms in their natural setting. (24). Historically microbial ecology was reductive based solely on the ability of microbes to be cultured, analyzed, and enumerated (25); leading many to think microbes meeting this criterion comprise the dominating members of the environments they were isolated from. However in most environments biodiversity is dominated by uncultured microorganisms. In some environments it is estimated that as many as 99% of the endogenous species are uncultivable with existing methodologies (26). The inability to cultivate may species is thought to be a result of experimental failure or a lack of knowledge of the real conditions under which most of the bacteria are growing in their natural environments (27). This means that surveys of microbial community members relying solely on culture-based techniques such as plate counts or most probable numbers (MPNs), will likely severely underestimate the extent of microbial diversity. Even “comprehensive” approaches such as standard plate counts or spectroscopy were aimed at studying important ecological variables such as total biomass, population sizes, fundamental processes and diversity of cultured organisms

then depending on the characterization of these cultured bacterial isolates to discern phylogeny (28). However, these methods lack the ability to link biomass, rate functions and diversity to fundamental contributions (12).

One of the original discoveries that piloted a true appreciation for microbial diversity was the use of direct counting methods by florescence miscopy (29), leading to the immediate realization that microbial biomass in natural environments was orders of magnitude higher than previously thought. Not only was the species composition greater, nearly all the cells within these biomasses were metabolically active and therefore contributing to community structure (30, 31).

The appreciation for unexplored diversity in microbial environments led to the idea that retrieving DNA from the total environment would, in principle, contain genetic information about nearly all the organisms in a community. Traditionally sequencing was done through clone based methods, where whole community DNA or PCR enriched portions were cloned into E. coli vectors. (32) These clones could be screened for functional properties (in the case of whole genome DNA). Plasmids from each clone could then be sequenced using plasmid specific primers by Sanger sequencing by synthesizing DNA on a single-stranded template through incorporation of random chain terminators (33). These terminators would generate a range of different fragment sizes corresponding to the terminators. Reactions were then run on a gel or capillary to identify the length of each fragment. Each base (G,A,T,C) would be run with the template having as a different terminator. By running the four base terminal reactions the gel or capillary image could then be translated into a DNA sequence (34). Diversity could then be assessed through non-parametric statistical

analysis. One of the first studies of a complex environment led to the idea that nearly 10,000 different taxa could be found in a small 100g soil sample (35). This estimate for numbers of taxa was orders of magnitude greater than those discovered through culture based methods. As methods for exploring microbial diversity has progressed over the last few decades, a consensus has developed around the idea that microbial communities are far more diverse than previously recognized.

1.2.1. DNA extraction method.

For molecular methods attention must first be taken to minimize errors and bias caused by impurities while processing environmental samples for their genetic content. DNA extraction is based on a series of extractions to remove substances such as bile salts, polysaccharides, urea, collagen, heme, myoglobin, hemoglobin, lactoferrin, calcium ions, or dnases, rnases, and proteinases (36) that can inhibit future molecular screening techniques (37). Accepted methods place sample materials in a tube containing beads and subject the samples to shear stress in attempt to break cells. Remaining material is subjected to an enzymatic cell lysis step to efficiently release cell contents into solution (including genetic material). Further purification removes PCR inhibitors, degrade protein, clean up the extract and finally concentrate the genetic material. Resulting extracts can then be used for downstream molecular testing. (16, 38-41). In a microbial survey, either total environmental DNA or an enriched fraction of the DNA obtained from the environment can be utilized.

1.2.2. Use of polymerase chain reaction (PCR).

The advent of PCR techniques (42) allowing for selective amplification specific DNA segments from complex mixtures of organism in an environment was an enabling step in development of metagenomics. Even when sample material is abundant such as fecal matter, soil, tissue, etc. contaminating materials from these environments can make DNA extraction problematic. However, PCR techniques allow for purity to sacrifice quantity of DNA recovery. Since only a small fraction of DNA is needed to achieve quality PCR reactions, DNA isolation techniques can focus on removing environmental contaminants, i.e. inhibitors.

PCR ultimately led to the ability to amplify genes across a wide variety of bacterial species. For the first time uncultivable or difficult to detect species could be monitored by DNA sequence analysis from a microbial environment. The adaptability of PCR and more specifically, the flexibility of PCR primer design, was soon realized leading to PCR primers targeting genes of individual species, conserved genes of functional significance, or even conserved genes from the majority of microorganisms. For example, specific primers are often used diagnostically to detect presence or absence of bacterial species such as pathogenic salmonella which can be differentiated by the presence of pagC/pagD (43). Primers targeting genes of specific functional groups can also be used to monitor microorganisms that share an important metabolic function such as butyrate formation in rumen bacteria (44) or the use of nifH in culture-independent studies of diazotrophs (45, 46). However, some situations primers are not designed for precise matches to known sequences. In cases of

targeting unknown or distantly related sequences for isolating genes encoding for known proteins degenerate primers are used. The strategy utilizes a pool of primers containing most or all of the possible nucleotide sequences discovered through a multiple alignment. Considerations should be made to avoid problems when studying highly divergent genes as increased primer mixtures and low annealing temperatures in these cases can cause artificial amplification. In these cases the gene of interest is often lost in the background noise especially when looking for highly divergent genes (47). A second approach in comparing distantly related species is utilized by understanding gene sequence of related organisms, consensus primers that choose the most common nucleotide at every position of multiple aligned nucleotide sequences, allow whole microbial surveys to be done. Using PCR primers that target conserved genes such as the 16S rRNA and rpoB or so called molecular clocks that are present in nearly all of the microbes in a certain environment and can be amplified and used in discerning phylogeny (32, 48-50). While useful for highly conserved gene homologs, or primer sites, mismatches between template and primer make consensus primers ill-suited for distantly related sequences. (47)

1.2.3. Development of a molecular clock.

The goal of a microbial assay was to tag a scientific name to a microbial isolate. Historically this was dependent on comparison of an accurate morphological and phenotypic description to that of a type strain or typical strain with that of the isolate in question. Microbiologists would utilize standard references such as Bergey's Manual of Systematic Bacteriology, the Manual of

Clinical Microbiology or form comparing isolates to well-characterized strains found in databases including the Centers for Disease Control and Prevention or the American Type Culture Collection that would summarize defining characteristics of each type species of bacteria (51-53). In many instances perfect matches were unable to be conclusively determined and conclusion would have to be made about the most probable identification. While these sources gave incite to help make ideal judgments identifications could easily vary among laboratories. (53) Further complicating this is that phenotypic methods can only be utilized on bacteria which can be isolated and cultured, limiting the methods to <1% of isolates (28).

It wasn't until 1980 that the work of Woese and others began to show phylogenetic relationships of all organisms could be inferred by utilizing and comparing a stable part of the genetic code (54, 55). Their work defined the 16S rRNA gene as the candidate gene for use in inferring microbial phylogeny. Its widespread use was adopted only after the 16S rRNA gene became categorized and its use for phylogeny was accepted across a broad range of microbial species. 16S rRNA gene sequence recovery from complex communities has permitted detection and phylogenetic assignment to microorganisms without the need of a preceding cultivation step (56). Currently universal primers have been applied with great success in virtually any environment within Earth's biosphere including (but not limited to) the mammalian gut, soil, and deep-sea subsurface sediments (16, 49, 50, 57). All prokaryotes contain complete ribosomes from genes responsible for coding the 5S, 16S, and 23S rRNA. From them the 16S rRNA has

become the leading gene of interest in determining the phylogenetic diversity of prokaryotes. The 16S rRNA is 1500bp nucleotide that assembles efficiently with the 30s small subunit (SSU) of prokaryotic ribosome. Effectively the 16S molecule underlies the structure and function of the ribonucleoprotein in prokaryotes (58). The 16S molecule forms a critical portion of the overall SSU and has the catalytic function of binding Shine-Dalgarno sequences of mRNAs to help imitate translation (59). Its usefulness in phylogenetic analysis is due to the fact that it has highly conserved variable and hyper variable regions. PCR products deigned to target the conserved regions will amplify DNA from a large population of prokaryotic cells and the amplicon spanning the conserved regions is variable enough to detect evolutionary shifts (54). These shifts can discern phylogeny for those same microbes allowing for further inference to provide a method of taxonomic identification (48). The evolution of the gene is congruent with speciation the 16S rRNA gene has been described as a 'molecular evolutionary clock'.

The 16S rRNA gene target is utilized through directed PCR at conserved regions which amplify either the entire 1500bp gene (60) or a segment of the gene containing one of the nine variable regions (61). 16S rRNA gene PCR primers can be designed to target a small subset of microorganisms or can be universal, amplifying nearly all eubacteria and archeal species (62, 63). The 16S rRNA gene was chosen due to its essential function in the host, occurrence in all prokaryotes, functional consistency, and the fact different positions in the sequence changes at different rates allowing the organisms to be studied across evolutionary time (64).

The gene can be easily extracted and sequenced directly and rapidly, and the 16S rRNA gene consists of many domains allowing them to have multiple sites for evolutionary comparisons (54).

Of course, any method suffers from limitations and bias. Problems with 16S rRNA arise from the fact that microbes contain variable numbers of rRNA gene copies. Even within a species there can be 2-5 fold difference in rRNA copy number between strains. This can cause variation in PCR amplification and difficulties in analysis and interpretation of 16S rRNA studies. In this sense, some isolates may be over- or under-represented when 16S rRNA genes are used in community analysis (65). The abundance of ribosomes in the environment should therefore be a species-dependent function of the number of individual cells and their growth rates to adjust for copy number differences. When looking at complex communities this should provide an estimate of each species to the entire protein synthesis capacity of the community (66).

1.3. Pre-sequencing molecular microbial methods.

Though methods for preparation of whole environmental microbial genetic mass or amplified PCR sequences (amplicons) were developed in the 1980's, DNA sequencing was laborious at that time. Thus, several analytical finger-printing techniques were developed to allow for relatively simple comparisons of microbial diversity between samples. Methods such as denaturation and temperature gradient gel electrophoresis (DGGE/TGGE) and terminal restriction fragment length polymorphism (T-RFLP) gave relatively quantitative means for comparing microbial

content of multiple samples (67-70). Probe based methods such as fluorescent in situ hybridization (FISH) and probe-based platforms such as phylochips utilize probes to visualize community interactions. Stable isotope probing (SIP) was developed using radioactive isotopes to probe microbial functions in the community.

Early Sanger based clone sequencing of all 16S amplicons were applied to around 250,000 16S rRNA genes across all environmental communities (71). In two larger example sanger based clone sequencing of all 16S amplicons covered nearly 11,831 bacterial and 1524 archaeal near-full-length, non-chimeric 16S rDNA communities in the human gastrointestinal (GI) tract (32) and only 21,752 16S rRNA isolates were recovered and sequenced from 111 diverse physical environments in studying global patterns in bacterial diversity (72). In 2005, with the introduction of massively parallel (so called Next-Gen) sequencing cost and effort in analyzing microbial communities quickly caused the demise in the use of traditional clone libraries as one sequencing run can yield more 16S rRNA reads (while in shorter length) than the collective efforts of all previous sanger clone based sequencing projects (71). From its origin, progresses in microbial ecology have been closely connected to technical and methodological developments.

1.3.1. Denaturation and temperature gradient gel electrophoresis – (DGGE/TGGE).

DGGE or TGGE rely on the principle that short segments of double stranded DNA will denature into ssDNA a temperature or denaturant concentration necessary to overcome the collection of individual hydrogen bonds. Gradient gels

take advantage of the fact that three hydrogen bonds in G/C base paring contribute more significantly hydrogen bonding in dsDNA than A/T base pairs which contribute two bonds. Therefore, DNA of a higher G/C content requires a greater denaturant or temperature and will subsequently migrate further on a denaturing gradient gel. Complete DNA separation of DGGE amplicons is avoided by incorporating an artificial GC clap at the end of the amplified molecule. As the amplicon migrates through the increasing denaturant gel it will form a largely ssDNA fragment with a dsDNA bridge of guanine and cytosine at one end, effectively stopping the fragment at its denaturation point. Each band on the gel is a measure of a specific organism and the bands collectively provide an indication of species diversity. Changes in banding pattern over time or treatment suggest how community structure is fluctuating in response to the variable. Gradient gels can be used to identify spices in a community through excision and sequencing individual bands of interest.

The technology expanded versatility over traditional methods, and as a molecular analytical tool, allowing for a complex microbial environment to be studied collectively (73). The major drawback is its limited ability to only resolve most abundant or top taxa in a community. In addition, the 16S rRNA segment that is analyzed and the denaturant concentration range have to be modified for different groups of microorganisms. Low abundance microbes fail to produce strong enough bands for visualization (27). Samples with high levels of diversity are difficult to resolve with gradient gels and species phylogenetic information is limited to bands that are able to be removed and sequenced. DGGE/TGGE is

especially useful when a microbial community is dominated by a few members, as time points or multiple samples can be represented in one gel image and easily visually compared through presence/absence or intensity of bands. (27, 73).

1.3.2. Terminal restriction fragment length polymorphism – (T-RFLP).

T-RFLP is based on PCR amplification of a target gene (16S rRNA) where one or both primers are fluorescently labeled at the 5' end. DNA from the environment is amplified with universal primers; the resulting amplicons are subjected to a restriction reaction normally using a four-cutter restriction enzyme (74). While the amplicons obtained from the microbial samples would show only limited variation in length, restriction sites may be found at very different sites within the target amplified gene. These difference are sequence specific and therefore potentially taxon specific. The sequence fragments are separated through the use of capillary or polyacrylamide electrophoresis in a DNA sequencer measuring different terminal fragments by a fluorescence detector. Only the terminal fragments that have been labeled are read and the resulting graph is a visualization of the florescent intensity versus size. (75) T-RFLP in this regard can be used with DNA to show complex microbial communities and produce fingerprints of the general microbial community composition (76). In theory each peak should correspond to a genetic variant (which can be probed to a curated database) and the intensity to its relative abundance giving a comprehensive picture of microbial diversity. This is not always true as often several different bacteria in a community might give a single peak when

restriction sites for a particular enzyme occur at the same position. While T-RFLP is reproducible (68) technical problems arise including incomplete digestion of environmental DNA and the formation of pseudo terminal restriction fragments (T-RFs) created from amplification of single stranded products, both would lead to an overestimation of diversity (77). Even without these biases, when applied to complex microbial communities each discrete peak may result in an unknown number of distinct species.

1.3.3. Quantitative PCR – (qPCR).

Quantitative real time PCR (q-PCR) is a florescent analytical method. It utilizes florescent binding of dsDNA in a sample. Differing from traditional PCR data is collected as the genes are amplified in 'real time'. The method identifies the number of cycles a sample takes to become linear before reaching saturation, with fewer cycles signifying a higher the amount of a bacterial species. q-PCR provides an extremely accurate quantification but is limited by the number of samples that can be directly tested as specific probes are designed for each bacteria of interest. q-PCR has been utilized to examine total microbial communities and the relative proportions of specific phylotypes within a number of unique environments (78-80). q-PCR can address metabolic potential of the microbial biomass by exploring specific biological functions utilizing probes targeting functional genes (81, 82). q-PCR is limited in that it requires extremely accurate controls for inferring cell mass or gene copy. It is advantageous when highly quantitative results on specific microbes or particular functional potential

of the community needs to be studied and prior knowledge of the sample is necessary for meaningful probes. Careful attention needs to be developed in producing secondary microbial specific assays to allow for creation of standard curves required for quantification. (83).

1.3.4. Stable isotope probing – (SIP).

Previous taxonomic and functional screens are limited in that not all microorganisms screened for a particular function will necessarily share the same genetic makeup as functional redundancy exists. A particular functional group may be underrepresented through the exclusion of these organisms and their function to the community. An alternative approach to link metabolic function to phylogenetic identity is to isolate microbes by their function and then determine their identity using molecular methods. Coupling molecular biological methods and staple-isotope probes in biomarkers, a cultivation independent strategy can be used to link bacteria with their environmental functions (84). SIP begins by introducing stable isotope-labeled nucleic acid or fatty acid substrates into a microbial community. DNA or fatty acid synthesis during microbial growth on a substrate enriched with a 'heavy' stable isotope (C^{13}) becomes labeled and can be resolved from unlabeled DNA by density-gradient equilibrium centrifugation in a CsCl gradient (85). Organisms that appear labeled are then associated with the biological processes from which the label was obtained. In the case of labeling nucleic acid, members that are actively growing can be identified. Organisms branded as utilizing the radioactive incorporated substrate can be resolved from a

density gradient centrifugation. DNA sequencing and fatty acid profiles can then be used to identify taxa in the labeled pool (86). A principal concern is determining if the target organisms will utilize a substrate in high enough levels to collect a sufficient proportion of 'heavy' stable-isotope, without degradation, to allow for collection of the enriched community DNA. SIP main drawback in molecular ecology is that communities cannot be studied in their entirety as functional screens must be developed for each community function of interest. On the other hand, SIP provides tremendous information about the relationships of taxa and function in different ecosystems or segments of an ecosystem.

1.3.5. Nucleic acid hybridization arrays

Hybridization arrays (microrarrays, beadarrays and phylochips) utilize specifically designed probes attached to a solid surface that can bind fluorescent dyed DNA complementary to those makers. The labeled DNA originating from the sample can then be assigned phylogeny or function based on the information known about the original probes. Arrays rely on oligonucleotide probes of 16S rRNA from specific groups of organisms to discern phylogeny. (87, 88) and, as such, can define community composition or function. Community genome microarrays would be useful in identifying core differences of critically important microbes in a sample. Through phylogenic, functional, and community genome microarrays are able to capture a much broader characterization of community structure than the above mentioned PCR based methods. This technology would be well suited for identification across multiple specific groups of microbes or a

gene functional group that is significantly important to the experiment or environment studied. However prior knowledge of the microbial composition is necessary for designing meaningful probes specific to the environment being sampled (87). In this regard microarrays often show highly skewed distributions of microbial species (89). The signal to noise ratio of array-based studies also suffers from cross hybridization between closely related species, genetic variations between strains within species and differential efficiencies of isolation DNA from a heterogeneous mixture of species also cause problems with the array technologies.

1.3.6. Fluorescent in situ hybridization – (FISH).

Unlike other nucleic acid-based assays, fluorescent in situ hybridization (FISH) allows visualization of the collective microbial community, such that phylogeny, morphology, localization, and abundance all can be measured. FISH is a probe based system comprising fluorescently labeled nucleic acid probes that bind to taxa of interest. These labeled oligonucleotide probes are diffused into fixed and permeable cells effectively labeling intact cells. . Florescent labels from a fixed environmental sample are then viewed under confocal laser scanning microscopy. Multiple hybridizations are possible because probes labeled with different fluorescence dies can be applied in parallel to the same hybridization experiment (90, 91). For taxonomy-based studies FISH utilizes oligonucleotide probes specific for the 16S rRNA, with the degree of conservation of the probe target sequence governing the level of taxonomic depth that can be discerned by

the probe (92). Alternatively, functional genes can be targeted, allowing visualization of all cells contributing to a function of interest.

FISH has been used to examine symbiotic interactions between viable cells within complex microbial environments and as a method to measure how interactions are effected by stimulus (90). Clearly, one of the greatest advantages of FISH is the direct visualization of labeled cells. However there are several reasons for the absence of FISH signals. The detection limit of FISH is high as between 10^3-10^4 cells/ml (or copies of the target sequence) of sample are needed for visualization (92). Targeted cells with a cell envelope may be less permeable to the fluorescent probes after standard formaldehyde fixation; as in many gram-positive bacterial requiring pre-treatment steps (93, 94). Target sites of rRNA may be inaccessible to the probe due to a high secondary or tertiary structure or binding of ribosomal proteins (95). With significant variation from permeability and target accessibility, one could well develop a biased view of the community. Thus, some prior knowledge of community structure is required to design probes encompassing critical areas of the community structure and make meaningful FISH experiments.

1.4. Sequencing.

Quantifying the degree of microbial diversity was hampered by the inconsistency between the degree of the microbial community that could be measured and the actual community size. The advent of high throughput massively parallel sequencing technologies allowed for more environmental samples to be sampled at a higher level

of phylogenetic diversity, generating more robust methods of inferences between environments. The goal of a metagenomic sequencing study is to sequence an entire microbial community without isolating or cultivating individual organisms. This is accomplished by sequencing the mixture of bacterial genomes comprising the community of DNA. The metagenome are the set of microbes that make up the environment together their genetic mass makes up the collective genome. As a comparison, if present in equal amounts a community of 1000 bacterial species would produce a metagenome roughly the size of one human genome (96-99). The taxa in a metagenome are not present in equal prevelence as the genomes of a small number of taxa may comprise greater than 90% of the biota while the genome of hundreds of rare community members can represent less than 1% of biota.

DNA sequencing can be used to quantify abundant community members and gain tremendous insight into phylogenetic and functional relationships. Early studies used Sanger sequencing (33) of 16S rRNA gene libraries made by cloning 16S rRNA amplicons into E. coli (32, 48, 55, 100). Sequence by clone based studies afforded the first high-resolution snapshots of microbial communities. Because of the cost and effort required, sequencing depth for each sample was limited, usually to the 100 clones. Nonetheless, tremendous strides were made in understanding community structure.

1.4.1. Massively parallel sequencing.

In 2005, a breakthrough in parallelizing DNA sequencing was announced, leading to the so-called “next-generation” of sequencing. With this technology

large communities can be studied based on phylogeny and/or function. Two platforms are commonly used for sequencing study.

1. The Roche 454 FLX platform uses “pyrosequencing.” This occurs through incorporations of a deoxynucleotide triphosphates (dNTPs) base into a synthesized DNA chain which releases a pyrophosphate. The pyrophosphate subsequently serves as a substrate for enzymatically production of ATP. In the presences of luciferase the ATP then leads to production of a quantifiable amount of light. The light output is then detected by a camera. This reaction is carried out on beads that contain millions of copies of a single DNA molecule. Two major parallelization steps were developed; the first was to simultaneously reproduce DNA fragments bound in a 1:1 ratio to small beads - These molecules are then colonially amplified in an oil-water emulsion containing PCR regains in micelles that are able to be occupied by only one bead. After harvesting, the beads are settled into wells of picotiterplate containing millions of wells along with enzymatic reagents. Sequencing occurs through repeated cycling flows of thin films of dNTP across the wells. Base incorporation leads to production of photons that are detected by a camera every seven seconds. Each cycle contains a different dNTP and a picture is taken after each cycle measuring light produced in each well. Thus the sequential collection of images are analyzed to measure the intensity of light. The amount of light which determines if a specific dNTP in that flow was incorporated or how many dNTPs were incorporated in that flow when homopolyer runs are present. This is then translated to DNA sequence for each bead. (99, 101)

2. The Illumina (Solexa) GAIIx platform functions by attachment of DNA fragments to a plate through hybridization of oligonucleotide adapters linking the DNA fragment to oligonucleotides on the plate. Fragments are amplified locally within clusters creating locally high densities for identical fragments. Flows of four different fluorescently dyed dNTPs are run over the plate. These highly specific dyed dNTPs distinguish the bases and block further incorporation once a dNTP has been added to the DNA fragment. After a round of synthesis a camera records the fluorescent signal, the dyes are cleaved (freeing the 3'-OH of the chain), and another cycle of reagents is able to be added and integrated. Images of all the fluorescent images are analyzed to create a DNA sequence. (99)

Both technologies analyze samples through a sequence by synthesis method. Software sorts out high quality sequences by sample which can be compared across multiple samples and runs. (102-107) Differences arise as the Illumina platform has much more sequencing depth at 40 Gb compared to Roche's 400Mb while Roche 454 produces long reads of greater than 400bp (106) lengths compared to Illumina which is currently producing reads up to 150bp (112).

With the ability to sequence hundreds of samples on a single sequencing run to depths previously unattainable by traditional clone based sequencing, the massive parallelization made it possible to characterize features of complex microbial communities in depth, providing detailed descriptions of composition previously unattainable.

Next generation sequencing has been utilized to describe and define community composition in multiple environments especially in cases were environments were under-sampled or their microbial composition was previously unknown. Comparative studies between these environments determine how diversity is distributed between samples. Quantitative results allow for the unique ability to understand shifts in microbial populations in an environment. The technology has made it possible to explore microbial communities in the quantitative interactions within them. Parallelized sequencers uncovered a much higher microbial diversity than previously understood. (16, 40, 41, 50, 108, 109). Two approaches are commonly utilized to characterize microbial communities: 16S rRNA gene sequencing (amplicon sequencing) defines community composition based on taxonomy of the microbes present in a sample. Whereas metagenomic whole genome sequencing (shotgun sequencing) utilizes the collective genetic mass, in this regard it is able to uncover phylogeny and microbial functions present in the environmental sample. However, much more effort, in the laboratory and computationally, is required to complete such a study compared to amplicon analysis, especially when dealing with a large number of samples.

1.4.1.1. Amplicon Sequencing

Ultra deep sequencing of target PCR amplified sequences from the 16S rRNA gene provides a relatively simple means to identify and quantify organisms in an environment. Total DNA can be isolated for hundreds of

samples, and the 16S rRNA composition of each sampled is represented by utilizing universal primers for the 16S rRNA in PCR reactions from each sample. Using bar-coded primers, the PCR products from each reaction can be pooled and sequenced in parallel. The diversity can be estimated from the sequence data using the proportions of individual sequence types as a proxy for the relative abundances of organisms in the sample. By producing sequence of each microbial variant in an environment diversity can be directly explored even in samples were no prior knowledge of the composition exists. In this regard sequencing has enabled a more comprehensive view into the diversity of organisms dominating an environmental habitat (50, 110, 111). Currently none of the current next generation sequencing platforms allow for full length coverage of the 16S rRNA gene so emphasis has been on identifying variable regions most useful in species identification (61). Because of the longer read lengths (106) achieved through Roche 454 based pyrosequencing it became the platform of choice in 16S rRNA analysis (20, 40, 41, 49, 50, 57). The longer reads yield more information about the 16S rRNA and thus give a more accurate classification. The typical goals of a 16S rRNA survey are defining and comparing communities on the basis of their microbial phylogeny or taxonomic content (61). Surveys can be achieved without an absolute account of all the species present. On a biological basis, enough resolution is needed to distinguish whether samples have similar or dissimilar taxonomic or phylogenetic content. 16S rRNA amplicon sequencing allows for comparisons to be performed accurately and is

especially useful in defining variability between environments (113, 114), between healthy or diseased states (20), identifying factors that explain variation between many biological samples (16, 115). Barcoding samples allows up to 200 samples to be sequenced in parallel achieving an average sequence depth of 5000 sequences/sample. Software then separates the samples and bins respective sequences within each sample (71). This allows for sequencing to achieve a higher dynamic range across more samples at a cheaper cost per sequence than previously developed Sanger clone methods (111).

Technical problems have been observed which are intrinsic to sequencing error rates of 16S rRNA genes (116). Overestimating of the rare biosphere is likely and attributed to pyrosequencing errors (117, 118). Systematic artifacts may lead to overestimating taxon abundance (119) and primer pairs used in the study greatly influence estimates of microbial community richness and evenness (120).

By removing traditional cloning from the methodology and directly sequencing the 16S rRNA gene from pooled DNA much of the biological bias associated with cloning is removed. However amplicon products are still subjected to the biases inherent to any PCR based experiment. Amplicon sequencing is limited to describing quantifying the species present and even with long reads of modern 454 sequencers many organisms cannot be accurately classified below the genus level (111). While studies have compared technical replicates of sequencing reads on a small scale (16), the

reproducibility of amplicon sequencing across a large number of biological replcates is still a question that still needs a definitive answer (121).

1.4.1.2. Shotgun Sequencing

A survey of 16S rRNA genes is useful in defining the community present, however, unless the microbes identified are well studied and well classified the 16S rRNA gene alone will not provide information on the role the organism has in the microbial community (122-125). By utilizing total environmental DNA and sequencing the entire genetic content of a mixed microbial population not only will phylogenetic markers such as 16S rRNA or *rpoB, recA* (126) be detected but also the sequence of many dominant genes in the environment can be extrapolated; allowing for the functional properties of the metagenome to be extrapolated for the environment studied. Use of total environmental DNA is achieved by random sheering of the total environmental DNA into short segments (400-800 bases) and sequencing the short segments to assign functional properties and biological processes fundamental to the microbial environment (127). Initially metagenomic work was done on the Roche 454 pyrosequencing platform (39) as short reads were too difficult to assemble and assign phylogeny accurately without known reference genomes, however, with length increases in the Illumina GAIIx platform from its initial production of 75bp reads to the 150bp now the reads generated do not appear to be a challenge for assembly of metagenomic whole genome sequencing. Furthermore assignment is streamlined with the use of

the increasing amount of reference genomes publically available. Currently, Illumina is the platform of choice as it can sequence much deeper, 40 Gb compared to 454's 400Mb, for a slightly higher cost. The short reads are sufficient for preforming BLAST searches and comparisons of sequence data to reference genomes. However, short reads seem to be limiting in assembling genomes from metagenomic studies. By using curated databases to assign function and since many of the genomes of the environment are from unknown organisms the technique probably underrepresent the true diversity within the environment. With the cost of sequencing dropping new reference genomes are constantly being sequenced; creating better databases of reference sequence which is critical to expand the information obtained from metagenomic studies (111). Approaches using highly curated de novo assemblies of complex metagenomes have also been successfully used to understand community structure and function by mapping reads or using BLAST to make informative sequences from other environmental samples to the original de novo assembly (17, 128). For exploratory purposes the technology would be ideal to search for possible microbial keystones within environments or for gene content that is key to a microbiome development (39, 113, 114). In metagenomic studies, overestimating diversity is likely as all reads have the potential to include sequencing errors intrinsic to next generation sequencing platforms (119). Therfore taxa and gene assignments can be skewed as these errors may lead to false assignments.

1.5. Bioinformatics

Sequence output from Next-generation platforms is substantial. Production of 16S rRNA sequences are 10-fold over previous clone based Sagner sequencing methods leading a million sequences produced in a typical 454 run. Typical metagenomic whole genome sequencing can produce terabytes of sequence.

While a true alignment based analysis is still the best method for analyses of sequence data, the depth produced carries too high of burden on current computers. To handle this volume of data new approaches in bioinformatics to discern phylogeny and community structure have been developed. Currently two broad types of analysis are usually employed to a break down samples: database dependent and database independent approaches. While far from exhaustive, the methods describe summaries of the common methodologies utilized in examining sequence data.

1.5.1. Database dependent bioinformatics

Amplicon database dependent analysis: Database dependent informatics is so deemed because classification of the query environmental sequence is dependent or matched to an entry previously curated in a database. 16S rRNA sequences from each sample in an experiment are analyzed through known databases, primarily Greengenes (129), SILVA (130) or Ribosomal Database Project (RDP) (131). The databases are utilized to assign a query sequence to a hierarchical taxonomy. RDP CLASSIFIER uses a naïve Bayesian rRNA classifier which is trained on known type strains of 16S rRNA sequences from its own curated database. To circumvent computation demands of alignment, frequencies of all

sixty-four thousand possible eight-base subsequences (words) are calculated for the training set sequences in each of the approximately 880 genera and the joint probability of observing the words in the submitted query can be calculated separately for each genus from the training set probability values (131). Following the naïve Bayesian assumption a query is assigned to genera with the highest probability (132). For analysis only a subset of the words are used for joint probability calculation, and the random selection and probability calculation is repeated for 100 trials. The number of times a genus is most likely out of the 100 bootstrap trials gives an estimate of the confidence in the assignment to that genus (61). Higher-order assignments sum the results for all genera under each taxon.

Both Greengenes and SILVA are utilized through an accessory program ARB(133). SILVA contains 618,442 high quality small subunit (SSU) bacterial isolates and Greengenes contains 1,049,116 SSU sequences over 1250bp. ARB takes these established databases and handles them using hierarchical taxonomy of the sequences and their associated information. ARB with SILVA or Greengenes creates phylogeny through suffix trees to find 40 closely related sequences creating a reference alignment. Reference sequences are transformed into partial-order graphs while still preserving their positional identity. This graph allows for swapping between different references to create optimal alignments (133). Further variability statistics are applied to give weight to conserved proportions and results are reported to the database in ARB. Alignment quality scores are given for each query sequence where sequence identity of over 90% is considered high classification. After taxonomic assignment, samples are

normalized by counts and statistical analysis is performed to probe any significant taxa between communities of interest.

Whole genome metagenomics database dependent analysis: When utilizing metagenomic whole genome sequence, databases are still utilized to assign function and phylogeny to sequence reads or to create a reference assembly. This is accomplished through comparison to reference genomes or to curated sequence databases. The first step in defining any metagenome involves a comparative analysis against various taxonomic and protein databases. This is done through tools such as BLAST which gives a snapshot of the community structure. Metagenomic assemblies can be done with software packages such as Roche Newbler, AMOS or MIRA. The software use reference data sets to compile metagenomic samples (134). In all reference based assemblies the quality depends on the availability, amount, and quality of representative genomes for comparison. Differences, such as insertions, deletions or polymorphisms, between the reference and sample can lead to fragmented assemblies or fail to cover divers regions (134).

These comparisons are at a high computational cost but provide the basic data for subsequent analysis such was phylogenetic profiling and comparisons, functional annotations, metabolic modeling and reconstruction, or simply binning of similar sequences (135). Further expansion of this basic study can be achieved by comparative analysis to annotated reference metagenomes or bacterial reference genomes. Metagenomic (MG)-RAST (135) can be utilized quickly to accomplish this task. MG-RAST unitizes the SEED framework (136) for

comparative genomes. With this system users upload relevant metagenomic data alongside raw sequencer output. Data is normalized, processed and summaries with annotations are automatically generated. The MG-RAST server then provides the users several tools to access different types of data from the ability to reconstruct phylogeny and metabolic structure to the capacity of comparing metabolism and annotations of one or more metagenomes or genomes of importance (137). MEta Genome ANalyzer (MEGAN) (138) also utilizes initial user generated BLAST output for a sequencing run. In this regard MEGAN is dependent on the NCBI taxonomy for phylogeny and NCBI Clusters of Orthologous Groups (COG) (139) classification for function. By performing a phylogeny MEGAN places each read of a given dataset into one taxa of the NCBI taxonomy. Reads matching significance in more than one BLAST species are assigned taxonomy to the lowest common ancestor (LDA) (138). Functional analysis utilizes the COG annotations which clusters genes into functionally related groups. Sequence is assigned into these COG categories by abundances. While COG analysis is readily incorporated in the BLAST input file used by MEGAN and is still used in publication, the COG classification is no longer curated, limiting its usefulness (140). To incorporate a more sophisticated functional analysis MEGAN also has the ability to use Gene Ontology (GO) (141) as a classification structure for binning environmentally generated sequence. GO annotations provide three hierarchical levels of ontologies. 1. Molecular function describes what the gene does at the molecular level. 2. Biological process describes what cellular processes the gene participates in. 3. Cellular component

details where the gene product is usually found within the cell (140). MEGAN again uses a LDA approach to assign each read to one node in each of the three GO ontologies. GO ontologies are not directly reported by BLAST so MEGAN uses a ref-seq table to assign GO terms to sequence reads. Reads are assigned a specificity score based on the number of annotated genes for a single GO term and its decedents compared to the total number of annotated genes. Once an annotation is complete MEGAN offers many visual comparisons of metagenomes as well as a chart based output, correlating nodes to read abundances for downstream use in statistical applications (140).

Regardless, if 16S rRNA metagenomes or whole genome metagenomes are utilized all methods described above rely on a previously available and highly curated dataset to be utilized for meaningful conclusions. However when exploring novel environments or when looking for rare members that may be under sampled in databases alternative database independent methods are needed.

1.5.2. Database independent bioinformatics

Amplicon database independent analysis: Database independent methods rely solely on the sequence data generated to compare sequences based entirely on their similarity to one another and do not rely upon on any databases for assignments. Cd-hit, UCLUST, BLASTCLUST and RDP pyrosequencing pipeline are a few of the main strategies for identifying sequence relationships by cluster analysis. Clustering a sequence in a dataset requires an all by all comparison, and is therefore very time consuming. BLASTCLUST (142) and methods using it

compute the all verse all similarity, and in doing so are very time consuming. These methods require much computational effort and are therefore not suited for clustering large sequence assemblies. The ribosomal database project (RDP) (131) developed a sequencing pipeline for taxonomic independent analysis. In this process sequences are aligned using the infernal aligner which will assign a similarity score based on aligned set of sequences. In this the RDP pipeline is not a true database independent approach however the database is not utilized in assignment of taxonomy as the above mentioned datasets. Alignments generated from infernal are then used in a complete linkage clustering algorithm. The algorithm bins sequences through a user defined threshold. The advantage to the RDP pipeline is that clustered sequence files can be directly used by several common tools to measure alpha and beta diversity indexes. However, RDP is still limited to clustering of 150,000 unique sequences so is still not well suited for large scale experiments. The RDP pipeline is limited to 16S rRNA sequences as the initial aligner is still reliant on a training set of ribosomal sequences. To handle these large volumes of sequence data and to avoid the need for training sets direct clustering approaches are employed. Cd-hit (143) uses a greedy incremental clustering algorithm method. Sequences are sorted in order of decreasing length, with the longest being representative of the first cluster. Each remaining sequence is compared to the representatives of existing clusters. If the similarity with any representative is above a given threshold it is grouped into that cluster, otherwise a new cluster is defined with the sequence as its representative. Clustering is established using *k-mers* by calculating the amount of identical

residues over a 100 residue window which relates to a given threshold. UCLUST (144) uses a global alignment of the query to a target that exceeds the user given identity threshold. Query sequences are processed in input order with the first sequence defined as the first representative *seed*. Each global sequence matching the seed is binned according to the identity threshold. If the query cannot be binned in the current seeds it becomes the seed of a new cluster. While allowing clustering of millions of sequences, these methods are limited in that additional steps are required to assign taxonomy or function to representative sequence. After clustering approaches are completed samples are again normalized by counts and statistical analysis can be performed to probe any significant differences between communities of interest.

Whole genome metagenomics database dependent analysis: De novo assemblies from short reads directly are still challenging. New techniques are beginning to overcome this obstacle, such as methods that conservation at the gene level (145), are promising especially bacteria or archaea where species have high-coding densities (146). Assembly tools, such as SOAP (147)or Velvet (148), are develop based on de Bruijn graphs in which the software looks for areas representing overlaps between sequences. Metagenomic de novo assemblies often demand a large computational load utilizing hundreds of gigabytes of memory and consuming days of computational time to complete analysis. Further complicating assemblies, microbial communities usually have great strain and species variation limiting the use of assemblies assuming clonal genomes. Metagenomic assemblers, such as MetaVelvet and Meta-IDBA (149), have been

developed recently to remove clonal assumptions that may lead to suppressions of contigs for heterogeneous taxa. The assemblers attempt to identify representative of related genomes from sub-graphing of the entire de Bruijn graph. Metagenomic sequences can be otherwise partitioned into species bins through the use of *k-mers*. Sub-graphs or bins are then resolved to build a consensus genome (134). De novo assemblies of complex metagenomes have also been successfully used to understand community structure and function by mapping reads or using BLAST to make informative sequences from other environmental samples to the original de novo assembly (17, 128). However, these all assemblies after complete are still reliant on database information to make meaningful conclusions, and implementation of complex assemblies for metagenomics is still in infancy. Accuracy of assemblers is difficult to assess as metagenomic data has no references for comparison. A highly curated (database-dependent) dataset for a diverse microbial community with known reference sequences will always be required (134).

1.6. Limitations of molecular methods.

Each step of a community analysis is subject to error or bias including cell lysis, DNA extraction and purification (150, 151), choice of primer (61), and PCR conditions. Some of these limitations can be detected, such as chimeric sequence formation and others can be minimized through optimization of the PCR conditions. However, it is unlikely that a single set of parameters will be optimal for all microbes making up a mixed sample (152). There are some inherent

complications with amplification. Universal primers, in theory, should amplify sequences from all microorganisms equally. In practice this is not the case as amplification is shown to introduce bias in a community analysis (153, 154). Although the 16S rRNA marker has been invaluable in developing phylogeny-based taxonomies and discovering novel microbial diversity, a single gene clearly doesn't represent all diversity within an ecosystem. In a species such as *E. coli* too strains may share identical 16S rRNA genes but vary in genomic content by as much as 100% (155), and effects of horizontal gene transfer upon assorted reproduction mask overall phylogenies (156). Thus, a number of genetic markers may need to be introduced to provide the resolution necessary to study microbial community structure (126).

In biological studies the goal is to break down the sample to the fundamental unit of biological classification. Classification is essential for describing, understanding, and comparing communities at different spatial and temporal levels. However the species concept as applied to microorganisms is highly controversial. Currently, a species is assigned to a common species if their reciprocal pairwise DNA re-association values are at least 70% in DNA-DNA hybridization experiments under standard conditions and there is a 5˚C or less ΔTm between their purified genomic DNA (157). This is level of hybridization is comparible to 97% sequence identity between 16S rRNA genes or 94% identity of total genome level. (158) While species concept does not translate to the eukaryotic community it has been applied extensively to the prokaryotic world. While massively parallel sequencing has made great strides in uncovering

microbial diversity, a major problem in analyzing diverse communities exists in that only the most abundant organisms are able to be visualized. Because of this, true biodiversity remains vastly underestimated.

1.7. Applications

The development and applications of "-omics" techniques able to gain information from direct isolation of DNA in microbial communities is leading to a paradigm shift in microbial ecology; the birth of a new golden age of microbiology is underway. In this 'age of bacteria', microbial ecology and the use of genetic and proteomic methods to examine microbes have commanded a renewed appreciation of the vital role played by microorganism to support and maintain life on Earth.

1.7.1. Microbial ecology and population biology

With nearly 10^{30} bacteria on this planet (159) unique biological niches and microbial species are being discovered as modern microbiology extends across all of the terrestrial biomass.

Molecular microbiology has exposed the details of how microbes interact in nature. In general microbes live as groups in communities, while some as a consortium of multiple microbes in complex societies with massive diversity (160) and others as near mono-culture units (161). Discoveries have uncovered a staggering, complex microbiota in nearly every environment on the planet. The microbial load on Earth is now thought to exceed in weight that of all other living things combined. It is estimated to account for nearly half carbon and over ninety

percent of nitrogen and phosphorous biomass on this planet (162, 163) critical to sustaining life. Microbes are responsible for creating a habitable climate as well as creating a sustainable soil conditions for agriculture (164, 165). They are shown to be nearly ubiquitous as they can be found in surprising niches from thermal ocean vents (166, 166, 167) to thriving in arsenic rich environments (168), expanding our understanding on how life can adapt and arise. Though classical microbial methods have provided deep insights to adaptations of single species, there is a lack of knowledge in understanding how complex consortia of microbes assemble.

Microbial ecology impacts much of the human condition from the expansion of pharmaceuticals, quality assurance in food productions, control of disease causing microbes in consumer goods, and throughout industrial applications (163). Microorganisms also play critical roles in host-microbe symbiotic interactions with implications toward diet, genetics, and lifestyle (16, 17, 39, 110). They are crucial for breaking down organic matter into useable substrates for uptake in eukaryotes (13). Microorganisms are used in a wide range of applications beneficial to the human condition. Without them the manufacture of vitamins, amino acids, enzymes, and growth supplements would be limited. Microbes are used as bio reactors in the production of pharmaceuticals.. Their use is critical in the manufacture of many foods, including fermented products (169, 170) and also in the degradation of waste and harmful chemicals into safe or even useful byproducts (171-173). Probiotics and prebiotics are aimed at selectively

incorporating microbes deemed to be beneficial to the host into consumables (174-177)

The rapid rate at which microbes can evolve ensures that no permanent solutions to agricultural, medical, or environmental problems they cause and technology will need to continue to evolve alongside microbial communities. However, the emergence of molecular methods has facilitated for the first time an expansive examination of microbial community structures and functions. The molecular tools now exist to examine an entire community's structure and functions using both low-and high-throughput techniques. The detailed information provided by these methodologies will enable informed decisions regarding resource management to be made. The following chapters give two unique examples of how microbial methods, especially massively parallel sequencing, can be applied to study diverse complex environments. Both studies primary emphasis is the compositional differences of microbial communities. As such, 454 pyrosequencing of the 16S rRNA amplicon has been the method of choice.

1.7.1.1. Example: Selective breeding for feed intake characteristics in mice resulting in unique gut microbial communities that contribute to feed intake phenotypes.

To investigate the relationship between gut microbiota and nutrient intake, this study capitalizes on a unique model of genetic selection in mouse lines bred for energy balance characteristics. In this model, breeding lines

were developed from an original four-way parental composite and selected for high (MH) or low (ML) heat loss through 25 generations along with an unselected control line (MC). The selection was repeated in three independent replicates. The MH lines have been well documented to demonstrate substantially higher levels of activity and feed intake in comparison to the other lines; a correlated response to selection. Nonetheless, the MH lines are characteristically leaner than the ML lines. Though several physiological characteristics have been well studied in these lines, the ability to study whether the gut microbiota also co-evolved with the host through selection has not. We used deep 16S rRNA pyrosequencing on of the MH and ML selection lines to test for changes in composition. Results showed that selective breeding for divergent feed intake traits shaped unique communities of gut microbes. Distinctive features of these line-specific communities are shared across independent replicates of the selection, implying that community change arose through changes in host genetic architecture and not to drift. Perturbation of the microbiota with antimicrobials shows a reduction in these microbial signatures while also narrowing the gap between the lines nutrient intake. Gnotobiotic transplant studies, confirm a direct link between the selection microbiome and feed intake characteristics, as introduction of host specific floras into germ-free recipients of a different genetic background led to feed intake characteristics of the MH and ML lines. Co-mingling animals excludes most of the environmental component of the correlated feed response to heat-loss. This model points toward heritability of both the feed

intake characteristic and gut phenotypes produced by selective evolutionary breeding and imply that the gut microbiota has a significant impact on the feed intake of mice.

1.7.1.2. Example: Use of pyrosequecing to identify new indicators of fecal contamination and temperature abuse in leafy greens.

Microbiological testing is a primary strategy used to assess safety of foods and beverages. Though quite sensitive, current microbiological methods rely on cultivation of targeted organisms in order to elevate the population and/or to allow visualization and enumeration. Reliance on these cultivation-dependent methods poses many limitations, one of the most significant being the very limited number of taxa that can be cultivated and enumerated and the actual correlation of target organisms with risk. Even methods which enumerate the entire cultivable microbial load are limited by cultivation bias and the inability to easily identify the organisms and show only limited correlation between absolute numbers and food safety/quality characteristics. To circumvent these problems, this project utilizes cultivation-independent, community DNA sequence-based methods as a means for safety and quality assessment of foods, specifically leafy greens. Pyrosequencing was used on a leafy green model (spinach) harvested from multiple environments. Samples from the phylloplane of spinach were compared with maize to determine if species specific

signatures existed in leaf epiphytes. Bovine feces were also introduced to confirm that spinach phylloplane flora was distinct from that of bovine feces. Spiking experiments were also carried out to test whether pyrosequencing could potentially differentiate contaminated on non-contaminated samples. The data shows compelling evidence that the microbiota of spinach is predictable; it is distinct from fecal microbiota; and predictable correlated changes occur under contamination. Thus, cultivation-independent, DNA sequence-based approaches is an alternative to culture-based microbiological testing.

2. SELECTIVE BREEDING FOR FEED INTAKE CHARACTERISTICS IN MICE RESULTS IN UNIQUE GUT MICROBIAL COMMUNITIES THAT CONTRIBUTE TO FEED INTAKE PHENOTYPES

2.1. ABSTRACT

Although abnormalities in the composition of gut microbiota have been associated with several types of complex diseases including obesity, the underlying cause-effect relationships remain poorly understood. To investigate causality between gut microbiota and nutrient intake, this study has exploited mouse lines developed through genetic selection for energy balance characteristics. In this model, breeding lines were developed from an original four-way parental composite and selected for maintenance of high (MH) or low (ML) heat loss through 25 generations along with an unselected control line (MC). The selection was repeated three independent times. The MH lines have been well documented for their characteristically higher levels of activity and feed intake in comparison to the other lines. Nonetheless, the MH lines are characteristically leaner than ML lines. Comparison of the gut microbiota across selection lines by 16S rRNA pyrosequencing showed that selective breeding for divergent feed intake traits shaped unique communities of gut microbes in the MH and ML lines. Distinctive microbial features of these line-specific communities are shared across independent repetitions of the selection, implying that elements of community change arose through changes in host genetic architecture and not to drift. Furthermore, perturbation of the microbiota with antimicrobials shows a reduction in these microbial signatures while also narrowing the gap between the lines feed intake. Gnotobiotic transplant studies which introduced MH and ML floras into germ-free

recipients of a different genetic background led to feed intake characteristics of the MH and ML lines in the conventionalized animals. Co-mingling animals excludes most of the environmental component of the correlated feed response to heat-loss. Collectively our data points toward co-evolution of host feed intake characteristic and gut microbiota phenotypes during selective breeding and demonstrate the significant impact host genetics can have on complex traits that include the gut microbiota and host physiological factors.

2.2. Introduction

Micro- and macro-organisms are habitually associated and interactions between the host and microbes shape distinct environments(178). Interactions are primarily dominated by microbes as they can outnumber host cells by many orders of magnitude (179). The collective microbiota can provide metabolic functions lacking from the host (122). Naturally occurring populations also interact with pathogenic species and can influence colonization outcomes (114) or health and disease states(18, 49). Interactions may allude to the importance of symbiotic or mutualistic relationship in community structure (40, 180-182). However, recently studies have questioned the stability of GI phylogeny (183). While thousands of bacterial species are thought to comprise the GI tract, only five phyla (20) comprise ninety-nine percent of the mammalian gut microbiota. Abnormalities in the relative proportions of these phyla are associated with health and performance of the host (18-20, 49, 184). Compositional aberrations which affect ratios of phyla *Bacteroidetes* and *Firmicutes* are specifically associated with predisposition to obesity in humans (184), and similar

compositional differences are observed in genetically obese mice. These compositional differences are believed to affect weight gain, in part, through their effects on the host's ability to harvest energy (39, 40, 49, 110, 184).

Mutualistic relationships of between gut microbial symbionts and their hosts are believed to have co-evolved through long processes which resulted in host-specific selective niche opportunities for colonization by microbial species that have adapted to these niches. The evolutionary outcomes of these relationships are remarkable, leading to adaptations such as herbivory, arising from the development of unique anatomical structures that produce niches for cellulose-degrading microorganisms (185). Residing within a host, microbes have likewise evolved features that have adapted them to these environments, such as the ability to selectively utilize the baroque oligosaccharides in human breast milk to colonize an infant's digestive tract and protect the gut from pathogen invasion (186). Some microbial innovations are even as sophisticated as the production of exopolysaccharides that contribute to the development of the mammalian mucosal immune system (19). Radiation and further refinement of these symbioses are reflected in the results of comparative studies of the microbiota from extant mammalian species, where correlations are shown between phylogenetic composition of the gut microbes and phylogenetic distances of their hosts (40, 187, 188).

While extreme disproportions of major taxonomic groups comprising the gut microbiota are clearly associated with complex diseases, the relationship between natural genetic variation to the development of dysbiosis, and associated susceptibility to complex disease is not understood. Within a host population,

individuality in gut microbiota is a pervasive phenomenon, easily observed by the large numbers of taxa that are sparsely distributed among individuals, with only a relatively small core of microbial taxa that are found among most individuals. Though a small number of taxa, this microbial core comprises the bulk of biomass in human and mouse models (16, 17).

In the mouse model, the relative abundance of these core organisms is controlled by complex combinations of host polygenes and environmental factors (16), meaning that host genetic variation can have measurable effects on microbiota composition. While host genetics is known to predispose individuals to disease, the ability of genetic variation to indirectly predispose individuals to disease through dysbiosis is not clear. Studies of single-gene effects in monogenic models have shown that null mutations in genes such as leptin (ob/ob) and toll-like receptors (TLR) result in dysbiosis and the resultant disproportions in microbiota composition by themselves can cause disease characteristics when transferred to germ-free hosts. (189, 190).

When viewed as a "trait", composition of the gut microbiome behaves as a collection of complex traits, affected by multiple environmental factors (chance, exposure, diet) as well as genetic characteristics of the host. Though substantial changes in microbiome composition can be observed in monogenic models (e.g. knock-out mice (40, 49)), natural genetic variation is polygenic. Given the importance of these symbiotic relationships in the developmental and metabolic functions of a host species, a question immediately arises about the interactions of genetic diversity within a host population, form and function of the gut microbiota, and health and disease states. It may be possible that genetic diversity within the host

population can lead to microbiome composition that ultimately predisposes individuals to disease.

A major hurdle to experimental study of these questions is the limited number of models that reflect the polygenic nature of many complex human diseases. Within human populations studies of twins generally support host genetic contributions, though some data are conflicting (15, 110), likely confounded by the high degree of genetic and environmental diversity in humans. Animal models have the advantage of being able to carefully control environmental and genetic factors but they suffer from lack of models that reflect the degree of genetic variation in outbred human populations. One way to test hypotheses related to genetics and complex traits is thought genetic selection. Genetic selection can amplify traits of interest from an original pool of genetic diversity, leading to the emergence of novel phenotypes and correlated responses to selection. Moreover, genetic selection lends itself to study the end-products of selection along with intermediates affording opportunities to identify critical steps and pathways of co-adaptation or co-evolutionary processes.

To understand relationships between host genetics, microbiota composition, and nutrient intake, we exploited a well-studied set of mouse lines developed through periods of repeated selective breeding for feed intake characteristics (191-193). In the study reported here two of three unique replicates the selection lines were sampled. From an original four-way parent composite, the MH (high maintenance) and ML (low maintenance) lines were developed through selective breeding for heat loss, a measure of energy that is metabolized but not stored. The physiological characteristics of these lines have been well-studied (191-195) providing an

outstanding opportunity to study the effects of breeding on microbiota composition and its interaction with physiological characteristics.

In all three repetitions of the selection, a substantial correlated response to selection was observed in feed intake, with the MH lines characteristically consuming 30-40% more feed per body weight than ML lines, but has lower body fat. The caloric intake is partially balanced by increased expenditures through heat loss and locomotor activity in the MH lines, but a significant proportion of the difference in the feed intake between these two lines remains unexplained. (191-193). Here we report on detailed studies of microbiome composition across two replicates of the MH, ML, and control lines.

2.3. MATERIALS AND METHODS.

2.3.1. Founder experimental animals and selection process.

Three replicates in three breeding lines were created from an original four-way parent composite (Harlan-ICR, Harlan-NIH, Charles River Swiss-Webster (CFW), and Charles River CF1) as described (192) . Using selection, breeding pairs were chosen on the basis of caloric heat loss (kcal•kg-.75•d-1) using gradient-layer, individual animal calorimeters from Thermonetics Corporation (San Diego, CA; model 0601-S, gradient-layer Seebeck envelope). For each of the experimental replicates, selective breeding for high (MH) or low (ML) maintenance of heat loss was carried out continuously through 16 generations along with an unselected contemporary line (MC) (191, 192). Significant

responses to selection were observed as early as after 15G selection as were correlated responses. These correlated responses included feed intake relative to body size ($g \cdot kg^{-0.75} \cdot d^{-1}$), which at 15G was manifested by the low-line mice consuming only 81% of the feed consumed by high-line mice at 8 to 11 weeks of age (191).

Selection pressure was halted after the initial sixteen generations. Effective population sizes were then expanded to 26 liters per line-replicate-generation to reduce the rate of inbreeding and preserve allelic contributions. Heat loss measurements and selections were restarted at generation 42 through generation 50 following the above protocol for MH, ML, and MC lines which led to resumption of divergence in traits between MH and ML across all three replicates (193). The lines have remained under relaxed selecting since, using 26 liters per line-replicate to preserve diversity.

For the experiments reported herein, mice from generations 58-65 were used. Mice were caged in groups by line and replicate within the same facility and feed intake was measured bi-weekly between 8 and 12wks of age. Ratios for feed intake were calculated by averaging the feed intake for all replicate mice within a line and comparing the total feed of MH to ML lines. Statistics on feed intake we analyzed by using a mixed model whiched included a fixed effect of line and random effects of replicate a a replicate*line interaction. Contrasts used to define Line differences: MH – ML = ‘Selection Effect’ and (MH + ML)/2 - MC = ‘Asymmetry’. Data analysis was carried out using the mixed procedure (SAS. *proc glimmix*)

2.3.2. Experimental animals, survey protocols, and sample collection

This study can be divided into major experiments testing for 1. line-replicate-generation differences in microbiota, 2. perturbation of feed intake by antibiotic feeding, 3. transfer of feed intake characteristics with fecal microbiota by gnotobiotic transplant, and 4. effects of intermingling of lines in caging environments through co-mingling of lines.

2.3.2.1. Line-replicate-generation experiment

From MH, ML, and MC lines sixteen mice were chosen randomly at generation 58 and generation 65 representing 25 generations of selection and 33 or 40 generations of relaxation respectively. A minimum of three fresh fecal pellets were collected from each animal and stored at -80^{0}C until use. Mice were caged by group within the same facility and feed intake was monitored bi-weekly for the generation 58 animals over 8 and 12wks of age.

2.3.2.2. Antibiotic perturbation animals

Twenty male MH, ML, and MC mice were randomly selected from each replicate from generation 66 (replicate 1 and 2) or generation 65 (replicate 3) for feed intake measurements. Mice were reared in individual cages starting at approximately 13 wk of age and given distilled water for 3 wk, followed by 4 wk of a distilled water and antimicrobial mixture (2 g/L streptomycin, 0.6 g/L metronidazole, and 0.35 g/L neomycin), and then returned to distilled water only for 4 wk. Body weight was obtained at the beginning of the study, just prior to starting the antimicrobial treatment, immediately after the

end of the treatment, and at the end of the experiment. Feed intake was measured weekly on individual mice as the difference between feed in and out per unit of body weight ($g/d*g^{-1}$) with a one week adjustment period given after initiation of the antimicrobial treatment and after returning to untreated water. At the end of each treatment period a minimum of three fresh fecal pellets were collected from four animals per line across the entire treatment and stored at -80^{0}C until use.

2.3.2.3. Transplant and gnotobiotic animals

Mice (Swiss Webster , 7 to 8 wk at the start) were acquired germ-free from Taconic Farms Inc. (Germantown, NY; model SWGF-F, SWGF-M). and reared under gnotobiotic conditions in two isolators as described (196, 197). Each isolator had 21 animals (11 males and 10 females), and mice were in cages of 2 to 5 animals. For 1 wk, feed intake was recorded. Then, mice in one isolator were colonized with a fecal slurry derived from a single MH mouse, by gavage with the donor fecal slurry, and mice in the other isolator were likewise colonized with a fecal slurry derived from a single ML mouse. Feed intake had been recorded on the MH and ML mice, and the donor MH mouse was consuming 73% more feed per BW than the donor ML mouse. Feed intake in the recipient mice was subsequently collected for 4 wk. The first week was considered an adjustment period, and feed intake per BW for weeks 2 through 4 were analyzed for the effect of colonization.

2.3.2.4. Co-mingled animals

To determine if differential feed intake characteristics could be due to physical segregation of the MH and ML lines (e.g. drift), a co-mingling study was performed in which 10 MH and 10 ML litters were randomly chosen from each replicate from G64 (replicate 1 and 2) or G63 (replicate 3) and co-housed in the same cages. After weaning, 2 mice from the MH litter were housed with 2 mice from the ML line for 8 wk (mixed rearing). The remaining littermates were reared, with no co-mingling between lines (like rearing). For each paired litter, one mixed reared and one like reared mouse was chosen from both MH and ML lines for antimicrobial treatment and feed intake measurements. Ten mice from the MC line were also randomly selected for further data collection. Statistical analysis of co-mingled to like reared animals followed a balanced design, over a large sample of 240 total animals. Data was analyzed using a mixed model. Replicate and Group through section of litters were random (a group is a pair of litters, one MH and one ML, that were co-mingled as well as reared within line), and all interactions of these two random effects with the fixed effects were also considered random. Fixed effects were Line and Type of Rearing and their interaction. The main variable analyzed was weekly feed intake (averaged for the 3 wk) divided by average body weight (feed intake relative to body weight).

2.3.3. Pyrosequencing.

A total of 16 mice from two replicate lines each of MH, ML, and MC were gathered at generation 58 (replicate 1 of each line) or generation 65 (replicate 2 of each line). DNA extraction from fecal pellets and pyrosequencing has been described previously (16, 41). The 16S rRNA was amplified from the DNA using barcoded fusion primers. The amplicons spans the V1-V2 region of the 16S rRNA gene and the fusion primers contain the Roche-454 A or B sequencing adapters (shown in italics), followed by a unique barcode sequence (N) and finally the 5' end of the 16S rRNA primer. Primers used in this study for FLX chemistry were A-8FM 5'-*GCCTCCCTCGCGCCATCAG*NNNNNNAGAGTTTGATCMTGGCTCAG-3‘ and B-357R 5'-*GCCTTGCCAGCCCGCTCAG*CTGCTGCCTYCCGTA-3. Primers used in titanium chemistry were A-8FM, 5'—*CCATCTCATCCCTGCGTGTCTCCGACTCAG*NNNNNNNNAGAGTTTGATCMTGGCTCAG and B-357R, 5'-*CCTATCCCCTGTGTGCCTTGGCAGTCTCAG*NNNNNNNNCTGCTGCCTYCCGTA—3’. PCR conditions (TaKaRa ExTaq) and pyrosequencing runs (454 Roche) follow the manufacturers’ recommendations. All PCR reactions were quality-controlled for amplicon saturation by gel electrophoresis; band intensity was quantified against standards using GeneTools (Syngene) software. For each region of a two-region picotiter Plate, amplicon reactions were pooled in equal amounts based on the GeneTools outputs to achieve ~8000 reads per sample and the resulting pooled sample was gel-purified (16).

Recovered products were quantified using picogreen ds DNA broad range assay (Invitrogen Q32850) by a Qubit fluorometer (Invitrogen Q32887), spectrophotometer (nano-drop ND-1000) and bioanalyzer (Agilent 2100) and sequenced using Roche-454 GS FLX/Titanium chemistry. Raw read output was filtered by length and quality procedures and binned by sample specific barcodes. For each region of a two-region picotiter plate, amplicons from up to 48 reactions were pooled in equal amounts.

2.3.4. Raw data filtering and binning.

Raw read data from the 454 pyrosequencing runs were processed through a quality filter removing sequences that fail to meet the following criteria:

- A complete forward primer and barcode present
- No more than 2 "N" characters (where N is equivalent to an interrupted and resumed signals from sequential flows)
- Length greater than or equal to 200nt but not longer than 500nt
- An average quality score above 20

After filtering, sequences were binned to sample-specific barcodes. Each read is trimmed to remove 3' adapter, primer sequences, and barcode. The corresponding FASTA and QUAL files were updated to remove quality scores from reads not passing quality filters. The files are associated with sample information in a hierarchical manner in MySQL tables. The processed data and the MySQL database tables are stored on a database server http://cage.unl.edu, allowing data to be made public after publication.

2.3.5. Taxonomic analysis and statistical method.

A CLASSIFIER+CD-HIT approach was subsequently used for determining taxonomic classification. Sequences were first parsed through the MULTI-CLASSIFIER (131) to separate those sequences that meet a genes-level threshold criterion. The algorithm assigns taxonomic status to each sequences read based on a covariance model developed from a training set of 16S rRNA sequences. Reads are classified down to the genus level at a threshold of >0.8 were further assigned species or OTU status using the best-BLAST hit to search a highly curated set of reference sequences from RDP CLASSIFIER (131) and the SILVA (198) databases. Sequences not meeting the BLAST criterion were assigned as an OTU with a genus level taxonomy. Sequences unable to be CLASSIFED at any taxonomic level were concatenated and binned through an OTU picking approach CD-HIT-EST (143). This algorithm groups related sequences on the basis of k-mer similarity by a cutoff threshold into OTUs for downstream analysis. A .97 sequence identity threshold and a .90 minimum coverage threshold were used to model 'species'- level phylotypes. As with classified sequences, the OTUs from CD-HIT were also used to query the curated 16S rRNA database to assign best-hit taxonomic status. OTUs or taxonomic bins were excluded if the bin failed to have on average 20 sequences per animal for taxonomic bins above this minimum, any animals having no counts were assigned 0.5 read abundance to preclude issues with log transformation.

After assigning taxonomic status through the CLASSIFIER/OTU pipeline, absolute proportions of communities meeting this criterion were conducted by:

$$absolute\ proportion = \frac{Number\ of\ reads\ in\ a\ taxon}{total\ number\ of\ reads\ in\ a\ sample}$$

Absolute proportions were log(10) transformed prior to statistical analysis and tested for significance (16). Significance testing was conducted between groups by ANOVA treating line and replicate as main effects for the replicate study. In the experiments with antimicrobial perturbations, treatment and line were the main effects. Scheffe testing was performed to check for line specific (replicate study) or treatment specific (antimicrobial study) differences. Associations between individual taxa and daily feed intake relative to body weight ($g/d*g^{-1}$) were tested by simple correlation analysis using Spearman rank correlation (199) for G58 animals between 8-12wk. All tests were deemed significant at $P<0.05$.

2.3.6. Antibiotic perturbation.

Twenty male MH, ML, and MC mice were randomly selected from each replicate at G66 (replicate 1 and 2) or G65 (replicate 3) and reared in individual cages starting at approximately 13 wk of age. The treatment regimen included distilled water (ad libitum) for 3 wk, followed by 4 wk of a distilled water and antimicrobial mixture (2 g/L streptomycin, 0.6 g/L metronidazole, and 0.35 g/L neomycin), and then returned to distilled water only for 4 wk. Body weight and

feed was monitored through the time course. In total, 16 mice, 4 from each line, were monitored over the entire antibiotic study and fecal samples were collected before (pre), during (on), and after (post) treatment and subjected to pyrosequencing. Pre-treatment body weight (PreBW), on-treatment feed per body weight (TrtFI/BW), and post-treatment feed per body weight (PostFI/BW) was analyzed with an as a mixed model with the fixed effect of treatment, line and the interaction of line*treatment. Random effects were for replicate only. Data was carried out using the mixed procedure (SAS *proc glimmix*) The main variable analyzed was weekly feed intake (averaged for each 4wk period of pre, on or post antibiotic treatment) divided by average BW, or feed intake relative to body weight.

2.3.7. Transplantation of microbiota into germ-free animals

Swiss Webster Mice were reared under germ-free conditions in two isolators as previously described (181). Each isolator had 21 animals (11 males and 10 females), with individual cages accommodating 2 to 5 animals. Mice were colonized by a ~100 µl gavage with the donor fecal slurry. All mice in a single isolator were colonized by fecal slurry from a single MH, or ML donor. MH and ML donors were chosen on the basis of feed intake measurements, with the MH donor mouse consuming 73% more feed per BW than the ML donor mouse. Feed intake and body weight of the gnotobiotic mice were monitored before and during the four weeks after colonization. Statistical analysis of pre-colonization BW (PreBW), pre-colonization feed per BW (PreFI/BW), and post-colonization feed

per BW (PostFI/BW) was done with a mixed model fitting treatment (MH or ML colonization) and sex (male or female). Preliminary analysis showed no effect (P > 0.95) of treatment by sex interaction on feed intake measures; hence the interaction was ignored in the final analysis. PostFI/BW also had PreFI/BW as a covariate in the model (SAS *proc GLM*) to adjust for pre-colonization differences.

2.4. Results

2.4.1. Effects of rearing (Co-mingling) on the feed intake trait.

Because the MH and ML lines were reared and housed by line, it was possible that any segregation of microbiome configurations could be due to drift and effects were due to caging within line. To test for this possibility, mice were co-mingled across line in the same cages. From the three independent selection replicates animals from ten MH and ten ML litters were caged together, two mice from an MH litter along with two from an ML litter. As a control, littermates of these animals were reared by line and caged three mice/cage. Any mixing of the microbial population between the MH and ML animals, combined with the role played by microbial population differences in feed intake, would therefore evince an interaction between Line and Type of Rearing. Under this hypothesis, one would expect co-mingling to result in less disparity in feed intake between MH and ML animals, statistically detectable as an interaction feed intake and caging. However, there is no evidence for the co-mingling interaction (p=0.30). The lines preserved their differences regardless of rearing in co-mingled or in like-line cages (Figure 1). Thus it can be concluded that the divergent gut microbiota

populations that are contributing to feed intake differences between the lines are not readily transferrable between lines after weaning; rather, they are due to the host genotype and the concomitant, early colonization events that occur pre-weaning. Line differences in feed intake adjusted for BW remain significant ($P<0.0001$). MH animals were shown to eat 40% more than ML at the same body weight characteristic of the trait in like-reared cages (Table 1).

2.4.2. Differentiation of the gut microbiota between the MH and ML selection lines.

The MH and ML lines were originally divergently selected for caloric expenditure though heat loss. Heat loss was used as the selecition criterion because it is an easily measurable phenotype which showed characteristically correlated responses in feed intake (191). The feed intake phenotype has been studied extensively as a correlated response in these lines (193) since this trait is of great interest to animal production. At the time of our study in G58 animals (25 generations of selection and 33 generations of no selection) the ML line consumes only 77% of the feed consumed by the MH line at 8-12 weeks of age. Previously, the lines have also shown differences in body composition but not body size: in comparison to the control, ML lines have higher fat content and MH lines are leaner (194). Collectively, studies of the physiological characteristics of these lines show that up to 80% of the phenotypic differences in feed intake between lines can be accounted for by metabolic characteristics, liver size, body fat, and

activity levels (193, 194). With 20% of the feed intake differences remaining unexplained, we began to test the hypothesis that selection shaped unique microbial communities that ultimately contribute to the line-specific feed intake characteristics. The dramatic phenotypic differences in dispensation of metabolized energy from feed between the MH and ML lines, the fact that these animals originated from the same composite base population, and that these lines have been in the same facility for >20 years therefore provided an excellent opportunity to examine the effect of evolutionary selection on gut microbiome composition.

The effect of selection on the compositional features of the microbiota was tested using deep pyrosequencing of 16S rDNA tags to estimate the relative abundances of bacterial species on total DNA extracted from feces from 12-week old MH, ML, and MC animals of selection replicate 1 (at G58) and replicate 2 (at G65). If the line-specific configurations of the microbiota were due to selection (and thus could be considered correlated phenotypes) as opposed to drift, similarities in the line-specific communities that were shaped across the independent selection replicates should emerge. The CLASSIFIER+CD-HIT pipeline identified eighty-four taxonomic bins having an average read count of >20/animal in the 96 animals sampled. These bins spanned six phyla, eleven classes, nine taxonomic orders, seventeen families, and fifteen genera. Probing these bins through BLAST produced twenty-six species classifications. These eighty-four bins represented the dominant members of the microbiome in these models comprise the principle taxa for our study (Table 2). ANOVA was

conducted across all of the principle taxa testing for main effects of Line and Replicate. Taxa deemed significant for replicate only or for mixed effects of Line*Replicate were discarded. This exposed only line specific taxonomic differences shared across both replicates which also behaved similarly across line and replicate. By these criteria, three families (Peptostreptococcaceae, Rikenellaceae, and Streptococcaceae), four genera (Erysipelotrichaceae Incertae Sedis, Lactococcus, Peptostrptococcaceae Incerteae Sedis, and Shegella), and three species (Two *Turicibactder* OTUs and *Lactobacillus apodemi*) were found with a significant line effect. On average a 0.84 absolute $\log_{10}$ difference was found between the MH and ML lines within these taxa across both replicates, with the minimum difference being 0.47 $\log_{10}$ and the maximum nearly 1.36 $\log_{10}$, with the members across these groups could comprised nearly seven percent of the an individual murine microbiota. To offset the error from multiple testing, post hoc Scheffe testing was conducted within taxa by line. All taxa found significant by line were also significant by Scheffe test between the MH and ML lines at $P<.05$.

The clear effect of breeding line on gut microbiota composition and the specificity of the effect across independent, replicate selection lines provide a compelling case for important compositional features of the gut microbiota as a correlated response to selection. Correlations between average body weights to daily feed in the G58 animals intake showed specific communities that provided a direct link between consumption, and microbial. While five phylogenetic levels representing fourteen significant taxon (Table 4) were shown to be significant as

correlated to feed intake in G58 animals, only three (genera Peptostreptococcaceae Incertae Sedis, species level OTUs *Turicibacter*, and species *Lactobacillus apodemi*) correlated with line specific results across both replicates, giving a higher likelihood of credence to their symbiotic role with line selection.

2.4.3. Antimicrobial modification of the line specific microbiota.

If the unique gut microbial communities were crafted as a correlated response to selection, it seems likely that the unique communities actually contribute to feed intake differences. As an initial test of this hypothesis, the microbiota was perturbed with large doses of antibiotic to determine if re-shaping the microbiota would affect feed intake. Following earlier studies on the relationship of microbiota and feed intake (200, 201), in our study mice were administered a combination of antimicrobials in their drinking water for a treatment of 4wks surrounded by 4wk periods without treatment (pre and post antimicrobials). Pyrosequencing of fecal pellets before, during and after antimicrobial treatment showed tremendous effects on the microbial community. Taxonomy-independent analysis of the gut microbiota from 4 pooled animals per line per treatment showed a threefold decrease in the number of OTUs were found pre-treatment than when antibiotics were administered (Figure 2A). All but three line-specific above taxa (genera *Shigella* and both species level *Turicibacter* OTUs) were affected significantly the antimicrobial cocktails ($P<0.01$) (Table 2). Thus, the antibiotic cocktail caused substantial changes in

composition of the microbiota including those taxa that could be contributing to feed intake. In an expanded study of 167 animals across three replicates, we monitored before, during and after antibiotic treatment. Feed intake, as divergence of MH and ML males in feed intake per unit metabolic size, was 41.8% pre-treatment and 28.3% while on treatment of the MC mean (Figure 2B). Antimicrobial treatments were shown to affect the feed intake in selection lines while also impacting taxa associated line specific microbiotas. This would imply that perturbation of the gut microbiota alone is sufficient to influence feed intake. Ideally, post-treatment the correlated response would return to pre-treatment levels after removal of antibiotics. However, this was not observed as a difference of only 17.8% was determined in feed intake between lines relative to the control. (Table 5). This small difference could be explained by the failure of the microbiota to return to pre-treatment diversity levels after being over-stressed with the antimicrobial cocktail as the pre-treatment had nearly 1.5 times as many OTUs as the suppressed post-treatment microbiota. (Figure 2A).

2.4.4. Transfer of gut microbiota from MH and ML lines to germ-free animals.

The antibiotic perturbation experiments had effects that are consistent with the hypothesis then selective for heat loss resulting in correlated responses to selection in the microbial communities, ultimately shaping communities that contribute to the feed intake phenotypes. To directly test this hypothesis, we next conducted a transfer experiment to determine if transfer of the MH or ML microbiota to naïve animals would also result in predicted feed intake

characteristics.Germ-free Swiss-Webster mice, a different genetic background than the composite from which the selections lines originated, were used as recipients in this experiment and prior to inoculation, were maintained in separate germ-free isolators (each isolator specific for MH or ML donor microbiotas). Mice in the two isolators that would be subsequently colonized did not differ in body weight before inoculation (PreBW) ($P > 0.95$); they did differ in PreBW by gender as expected ($P < 0.02$; males = 29.2 and females = 25.4 g, se = 0.9). No differences were observed in feed before inoculation (PreFI/BW) for either isolator ($P > 0.25$) or sex ($P > 0.35$). However, feed intake after transplanting the microbiome (PostFI/BW), adjusted for PreFI/BW, was different for mice colonized with MH microbiota as compared to mice colonized with ML (MH = 0.136 and ML = 0.126, $g/d*g^{-1}$, se = 0.003, $P>0.031$). While decreased feed intake of animals was observed in both isolators when the mice were colonized with microbiota, those colonized with MH microbiota had smaller reduction (Table 6). The initial decline in feed intake post colonization is typical in germ-free mice (202-205). However, these previously novice mice expressed 8% greater decrease in feed intake per BW with the ML colonization than with the MH. (Figure 3). These results therefore demonstrate that a portion of the line-specific feed intake characteristics can be transferred though the microbiota, further supporting our hypotheses that selection for heat loss led to unique microbial communities in the MH and ML lines and these communities contributed to the feed intake (and perhaps even the heat loss) phenotype.

2.5. Discussion

Comparative physiological studies of complex traits such as diet and gut microbiota composition can yield tremendous insights into the evolutionary patterns of complex traits in extant populations (40, 187, 188). Investigative approaches such as genetic selection experiments can provide a powerful means for understanding the mechanistic factors that are ultimately shaped by evolutionary processes (206). Previous QTL analysis showed that composition of the microbiota is heritable, that many of the individual QTL had significant additive effects, and that several of the individual taxonomic groups were controlled by multiple loci (16). Additive effects from multiple loci are a prerequisite for a complex trait to respond to selection (206) and the gut microbiota as a complex trait would have the requisite features to respond to selection. Our results with the MH and ML lines support this idea. Though we were limited to studying the end products of selection (the methods we used for microbiome analysis did not exist prior to 2005) the collective sets of analytical and functional experiments are all consistent with this conclusion.

Though it might be argued that line-specific caging led to differences in the microbiota, we note that mice originated from the same composite base population, were housed in the same facility, and shared the same diet. Co-mingling experiments further showed a negligible effect of the caging environment on the phenotype. While development of complex phenotypes can occur relatively quickly through artificial selection, the genomic changes model natural evolutionary events. Such models can therefore provide tremendous insight into the trait because intermediates can be studied (206). Thus, with respect to the microbiota, studying the assemblies of

organisms at differ stages of selection may reveal pathways through which host genotypes can translate into microbiome configurations. Selection models also make it possible to detect the changes in host genetic architecture and gut microbial communities which resemble the magnitude of an evolutionary event. Observation across multiple selection experiments often results in different pathways to a complex trait, even when selection is imposed on multiple populations of genetically similar (if not identical) individuals. This is particularly true when a trait is manifest by a combination of two or more components that also vary quantitatively (e.g. speed and duration in running traits) (207) and is likely due to the fact that all possible configurations of genomic architecture that underlie variation in the trait have not been sampled.

Though we did observe this phenomenon in the independent replicates of the MH and ML lines, significant microbial "traits" were found to have emerged in each replicate. For example, the individual species of *Lactobacillus apodemi* was on average 24.7% (Rep1-13.2%, Rep2-36.3%) higher in the ML over MH lines (Figure 4a).

As feed intake has been previously shown be associated with composition of the intestinal flora (110) it is not surprising that the microbiota responded to selection (Table 3). For example, in this study *L. apodemi* shown to be higher in ML lines, naturally, negatively correlated with feed intake (Figure 4a). There were three families (Peptostreptococcaceae, Rikenellaceae, and Streptococcaceae), four genera (Erysipelotrichaceae Incertae Sedis, Lactococcus, Peptostrptococcaceae Incerteae Sedis, and Shegella), and three species (Two Turicibactder OTUs and *Lacobacillus*

apodemi). The similar responses and the ability to transfer phenotype with microbiota transplant argue that these taxa are directly contributing to feed intake differences. There were three families (Peptostreptococcaceae, Rikenellaceae, and Streptococcaceae), three genera (Erysipelotrichaceae Incertae Sedis, Lactococcus, Peptostrptococcaceae Incerteae Sedis), and one species (*Lacobacillus apodemi*) taxa were also reduced upon antimicrobial treatment, which also reduced the line effect on feed intake. Interestingly *L. apodemi* was the only species level to show correlation with feed intake, to show line effects, and to show effects upon antibiotic treatment (Figure 4a,b). It is interesting to note that these particular organisms produces a tanninase (208, 209) The organism is known to produce gallic acid from tannic acid, but does not convert gallic acid further to pyrogallol (209). Tannins are phenolics found in all classes of vascular plants, categorized as secondary compounds produced by plants as a defense against herbivory. Tannins have the ability to precipitate proteins and in that sense can possibly interfere with digestion or the assimilation process (210).. In food animals tannins decrease intake and digestibility of dry matter and protein (211-214). It is therefore possible that specific gut organisms could play an essential role in digesting tannin containing foods allowing more caloric value of foodstuff. This could explain *L. apodemi* populations being higher in the ML lines possibly signifying how the MH line characteristically taking in more feed while remaining leaner as they would fail to acquire and absorb the extra caloric content available to the ML lines which contain *L. apodemi*.

To solidify the contribution of the microbiota to feed intake when germ-free recipients of an entirely different genetic background were transplanted with line

specific microbiotas, the GI inhabitants of the MH and ML lines were able to confer predictable, line-specific differences in feed intake. The dissimilar genetic background of the recipient further underscores the ability of the community itself to confer a measurable phenotype.

Of course, one of the most significant questions remaining is how genetic selection affected the microbiota. Mapping studies in these lines should prove fruitful and lead to loci indicative of the selection. Such studies would be particularly useful for pinpointing pleiotropic loci that influence both *L. apodemi* colonization and feed intake.

Beyond the traits studied here, this work also paves the way for using selection as a means for studying microbiota assembly and the evolutionary steps that contribute, without inference. In addition, the boundaries of microbiota composition can also be studied to determine if disease susceptibility is a correlated response to selection for extremes of microbiome composition (18).

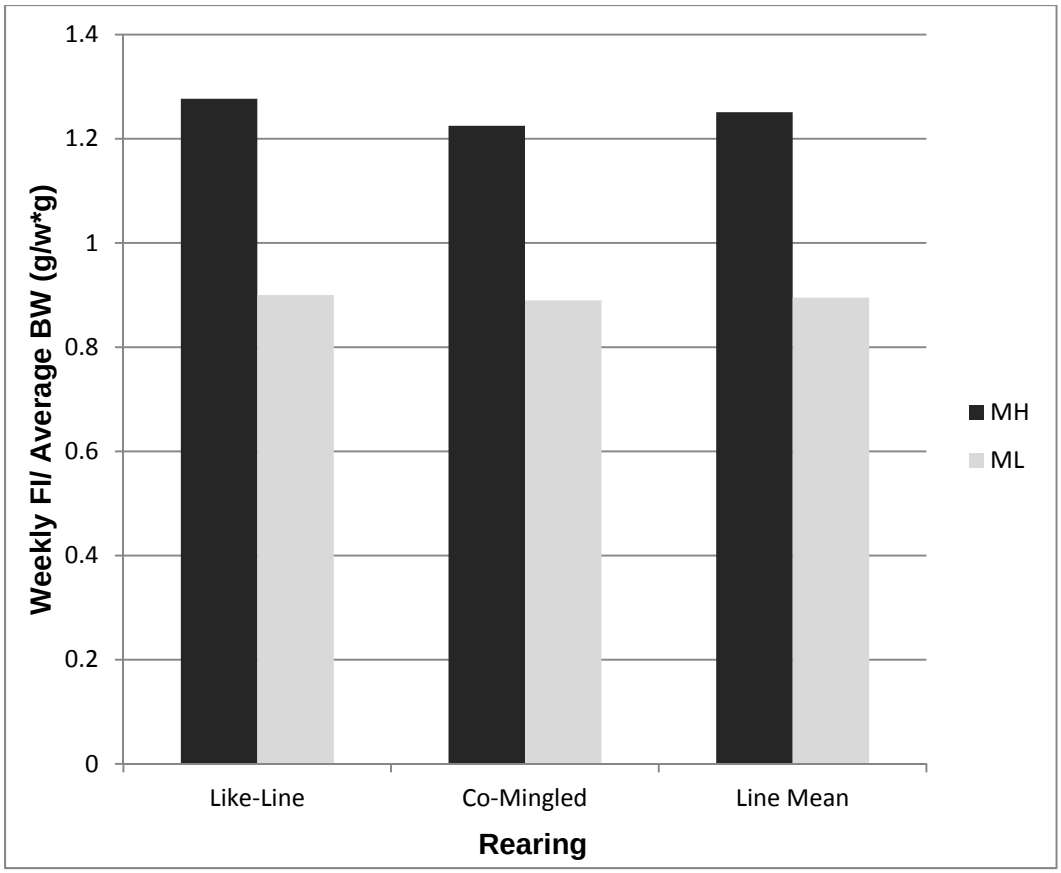

Figure 1: Effects of caging environments on feed intake of animals selected for heat loss characteristics. 240 mice across three repetitions were reared in duplicate cages of 10 mice for two rearing types (Like-Line or Co- Mingled) across the two lines (MH, ML). Graphs depict weekly feed intake per body weight (FI/BW) across caging for line. FI/BW was reported in grams bodyweight per daily grams of feed uptake (g/d*g). Line effects were averaged for Like-Line (no co-mingling) or co-mingled cages. Line Means were calculated from all 240 mice.

Co-Mingle Study	Like-Line	Co-Mingled	Line Mean
MH	1.277	1.225	1.251
ML	0.9	0.89	0.895
Rearing Mean	1.088	1.057	

Table 1: Feed intake per BW of animals reared in like lines or co-mingled across lines. Weekly feed intake per average body weight (g/g*w) across the rearing lines. Line means are the average across all animals regardless of rearing, rearing means are averaged across all rearing regardless of line.

Table 2: Descriptive statistics for principle taxa (3 pages). Principle taxa for this study were defined as having an average of >20 sequencing reads/animal at a taxonomic level. For these taxa descriptive statistics were calculated for the log transformed relative abundance across all animals for average, standard deviation, maxiumim, and minimum.

		Average	St. Dev.	Min.	Max.
phylum	Actinobacteria	-2.39434	0.590728	-4.4679	-1.36208
phylum	Bacteroidetes	-0.57927	0.320404	-2.41308	-0.02903
phylum	Cyanobacteria	-4.03672	0.569818	-4.61138	-0.93432
phylum	Firmicutes	-0.36521	0.261746	-1.27818	-0.02073
phylum	Proteobacteria	-1.68881	0.487336	-2.9502	-0.49166
phylum	TM7	-3.39097	0.859262	-4.4843	-1.55595
class	Actinobacteria	-2.39434	0.590728	-4.4679	-1.36208
class	Alphaproteobacteria	-3.32068	0.890751	-4.4843	-0.52972
class	Bacilli	-0.68788	0.46915	-2.05806	-0.06958
class	Bacteroidetes	-0.93381	0.464485	-3.00154	-0.30837
class	Betaproteobacteria	-2.86411	0.75026	-4.38306	-1.10372
class	Clostridia	-1.03787	0.46417	-2.62411	-0.2143
class	Cyanobacteria	-4.03672	0.569818	-4.61138	-0.93432
class	Deltaproteobacteria	-2.87234	0.825952	-4.44898	-1.05377
class	Epsilonproteobacteria	-2.2836	0.628382	-4.25681	-1.04379
class	Erysipelotrichi	-2.24675	0.847652	-4.34908	-0.48473
class	Gammaproteobacteria	-3.29548	0.856542	-4.4843	-0.69971
order	Bacteroidales	-0.93381	0.464485	-3.00154	-0.30837
order	Campylobacterales	-2.2836	0.628382	-4.25681	-1.04379
order	Clostridiales	-1.04768	0.46539	-2.62411	-0.21728
order	Coriobacteriales	-2.49237	0.614541	-4.4679	-1.41917
order	Enterobacteriales	-3.63545	0.771775	-4.61138	-0.70134
order	Erysipelotrichales	-2.24675	0.847652	-4.34908	-0.48473
order	Lactobacillales	-0.69272	0.475775	-2.09493	-0.07028
order	Rhizobiales	-3.80176	0.762639	-4.61138	-0.89056
order	Sphingomonadales	-3.63818	0.868106	-4.4843	-0.61882
family	Bacteroidaceae	-1.55493	0.514569	-3.35372	-0.51267
family	Chloroplast	-4.03672	0.569805	-4.61138	-0.93455
family	Clostridiaceae	-3.74356	0.838473	-4.51319	-1.15473
family	Coriobacteriaceae	-2.49237	0.614541	-4.4679	-1.41917
family	Enterobacteriaceae	-3.63545	0.771775	-4.61138	-0.70134
family	Erysipelotrichaceae	-2.24675	0.847652	-4.34908	-0.48473
family	Helicobacteraceae	-2.28361	0.62837	-4.25681	-1.04379
family	Lachnospiraceae	-1.33334	0.499183	-2.87938	-0.35874
family	Lactobacillaceae	-0.71609	0.49481	-2.16004	-0.07123
family	Peptostreptococcaceae	-3.65407	0.86154	-4.51319	-1.08144
family	Porphyromonadaceae	-2.38454	0.626566	-4.38306	-0.85855
family	Prevotellaceae	-3.06415	0.730362	-4.44898	-1.57061

family	Rhizobiaceae	-4.14959	0.48077	-4.61138	-1.02605
family	Rikenellaceae	-1.83064	0.75334	-4.4679	-0.8258
family	Ruminococcaceae	-2.32499	0.67136	-4.4679	-0.54921
family	Sphingomonadaceae	-3.63818	0.868106	-4.4843	-0.61882
family	Streptococcaceae	-3.12621	0.785462	-4.3404	-1.21721
genus	Alistipes	-2.05877	0.677217	-4.4679	-0.96967
genus	Bacteroides	-1.5551	0.514532	-3.35372	-0.51302
genus	Clostridium	-3.76832	0.823659	-4.51319	-1.16601
genus	Erysipelotrichaceae Incertae Sedis	-3.17016	0.94474	-4.61138	-1.15754
genus	Helicobacter	-2.29416	0.64559	-4.28995	-1.04426
genus	Lactobacillus	-0.71942	0.495191	-2.16648	-0.07759
genus	Lactococcus	-3.31392	0.829018	-4.41377	-1.22412
genus	Parabacteroides	-2.40083	0.634747	-4.38306	-0.85941
genus	Peptostreptococcaceae Incertae Sedis	-3.66394	0.858528	-4.51319	-1.10533
genus	Ruminococcus	-3.61914	0.843662	-4.61138	-0.68661
genus	Shigella	-3.88223	0.624044	-4.61138	-1.02826
genus	Sphingomonas	-3.70748	0.857334	-4.4843	-0.62195
genus	Streptophyta	-4.03675	0.56963	-4.61138	-0.93616
genus	TM7_genera_incertae_sedis	-3.39097	0.859262	-4.4843	-1.55595
genus	Turicibacter	-3.13354	1.187263	-4.51319	-0.4876
species	Clostridium_OTU9	-3.78151	0.813841	-4.51319	-1.22125
species	Helicobacter ganmani (T)	-3.55526	0.795147	-4.4843	-1.15541
species	Streptophyta_OTU16	-4.0946	0.539671	-4.61138	-0.99179
species	Turicibacter_OTU70	-3.73477	0.812387	-4.61138	-1.22771
species	Turicibacter_OTU93	-3.75805	0.900007	-4.51319	-1.28669
species	Parasutterella_OTU11	-3.21623	0.849808	-4.61138	-1.14446
species	Sphingomonas oligophenolica (T)	-3.83808	0.824876	-4.61138	-0.68217
species	Odoribacter_OTU2	-3.09598	0.785764	-4.51319	-0.99962
species	Bacteroides_OTU20	-3.28093	1.042351	-4.61138	-1.12534
species	Oscillibacter_OTU2	-2.92827	0.705161	-4.4679	-1.22835
species	Alistipes_OTU3	-2.9088	0.763947	-4.4679	-1.3051
species	Bacteroides_OTU13	-3.07506	0.878459	-4.61138	-1.31745
species	Lactobacillus crispatus (T)	-3.77463	0.776013	-4.61138	-0.56819
species	Parabacteroides_OTU3	-2.62447	0.695518	-4.38306	-1.04925
species	Alistipes_OTU8	-2.57975	0.681052	-4.4679	-1.45364
species	Helicobacter aurati (T)	-2.59424	0.849658	-4.61138	-1.28214
species	Alistipes_OTU9	-2.45703	0.689239	-4.4679	-1.16105
species	Turicibacter_OTU49	-3.43598	1.03802	-4.51319	-0.62891
species	Lactobacillus reuteri DSM 20016	-2.43726	0.796834	-4.4843	-0.89584
species	Bacteroides_OTU0	-3.25729	0.849859	-4.37014	-0.53155
species	Bacteroides_OTU16	-2.25844	0.713655	-4.51319	-0.59828
species	Bacteroides_OTU5	-2.17711	0.677441	-4.18616	-1.0009
species	Lactobacillus reuteri JCM 1112	-1.67747	0.667279	-3.91113	-0.33262

species	Lactobacillus intestinalis (T)	-1.87164	1.021584	-4.4679	-0.45182
species	Lactobacillus apodemi (T)	-1.62747	0.766086	-3.7328	-0.23111
species	Lactobacillus johnsonii ATCC 33200	-1.55903	0.809901	-3.7385	-0.11077

Line Differences Across Two Repetitions		Replicate 1	Replicate 2	Replicate 1	Replicate 2	Replicate 1	Replicate 2
Taxonomic Rank	Taxa	C	C	H	H	L	L
family	**Peptostreptococcaceae***	-3.73 ± 0.67	-4.15 ± 0.19	-4.02 ± 0.40	-4.34 ± 0.12	-2.78 ± 1.04	-2.90 ± 0.82
family	**Rikenellaceae***	-2.29 ± 0.49	-1.33 ± 0.28	-2.51 ± 0.69	-1.67 ± 0.88	-2.00 ± 0.73	-1.19 ± 0.18
family	**Streptococcaceae***	-2.46 ± 0.66	-3.38 ± 0.78	-2.74 ± 0.85	-3.00 ± 0.59	-3.31 ± 0.54	-3.87 ± 0.44
genus	**Erysipelotrichaceae Incertae Sedis***	-2.40 ± 0.68	-3.54 ± 0.79	-2.57 ± 0.45	-3.50 ± 1.02	-2.80 ± 0.87	-4.21 ± 0.28
genus	**Lactococcus***	-2.69 ± 0.84	-3.56 ± 0.83	-2.85 ± 0.92	-3.19 ± 0.63	-3.61 ± 0.54	-3.98 ± 0.41
genus	**Peptostreptococcaceae Incertae Sedis***	-3.73 ± 0.67	-4.17 ± 0.17	-4.04 ± 0.39	-4.34 ± 0.12	-2.79 ± 1.03	-2.91 ± 0.82
genus	**Shigella**	-4.03 ± 0.51	-3.75 ± 0.63	-3.82 ± 0.57	-3.46 ± 0.98	-4.04 ± 0.33	-4.19 ± 0.25
species	**Turicibacter_OTU70**	-3.16 ± 1.17	-4.04 ± 0.31	-4.03 ± 0.27	-4.34 ± 0.12	-3.07 ± 0.98	-3.77 ± 0.56
species	**Turicibacter_OTU49**	-3.00 ± 1.28	-3.82 ± 0.73	-3.91 ± 0.41	-4.34 ± 0.12	-2.82 ± 1.33	-2.72 ± 0.61
species	**Lactobacillus apodemi (T)***	-1.52 ± 0.97	-1.60 ± 0.70	-1.83 ± 0.60	-2.15 ± 0.53	-1.39 ± 0.82	-1.27 ± 0.68

Table 3: Taxonomic differences between lines across two different independent replications selected for heat loss characteristics. ~8000 filter passed pyrosequencing reads were utilized for 16 unique animals across two distinct replicates at two different time points (G58, G65). The table represents significant differences by ANOVA across Line at p<.05. Three families, four genera, and three species were found to make up a significant line effect. (*) Represent taxa responding to treatment in the antibiotic selection model at P<.01.

Correlation Values on Feed Intake		Correlation Matrix	F Statistic	Correlation Significance
Taxonomic Rank	Assignment	R	F	P
Class	**Erysipelotrichi**	0.341	6.063	0.018
Order	**Erysipelotrichales**	0.341	6.063	0.018
Family	**Clostridiaceae**	0.595	25.177	0.000
Family	**Erysipelotrichaceae**	0.341	6.063	0.018
Family	**Peptostreptococcaceae**	0.508	15.983	0.000
Genus	**Clostridium**	0.594	25.044	0.000
Genus	**Peptostreptococcaceae Incertae Sedis**	0.527	17.645	0.000
Genus	**Turicibacter**	0.425	10.126	0.003
Species	**Clostridium_OTU9**	0.594	25.033	0.000
Species	**Lactobacillus apodemi (T)**	0.321	5.269	0.026
Species	**Lactobacillus reuteri JCM 1112**	0.329	5.577	0.022
Species	**Turicibacter_OTU49**	0.437	10.830	0.002
Species	**Turicibacter_OTU70**	0.380	7.778	0.008

Table 4: Taxonomic correlations across one replicate for feed intake characteristics. G58 animals were monitored for their feed intake between 8 and 12 weeks. Correlations were completed between FI/BW (g/d*g) and the log(relative abundance) of filter passed taxa. Significance was determined at correlations $p<.05$.

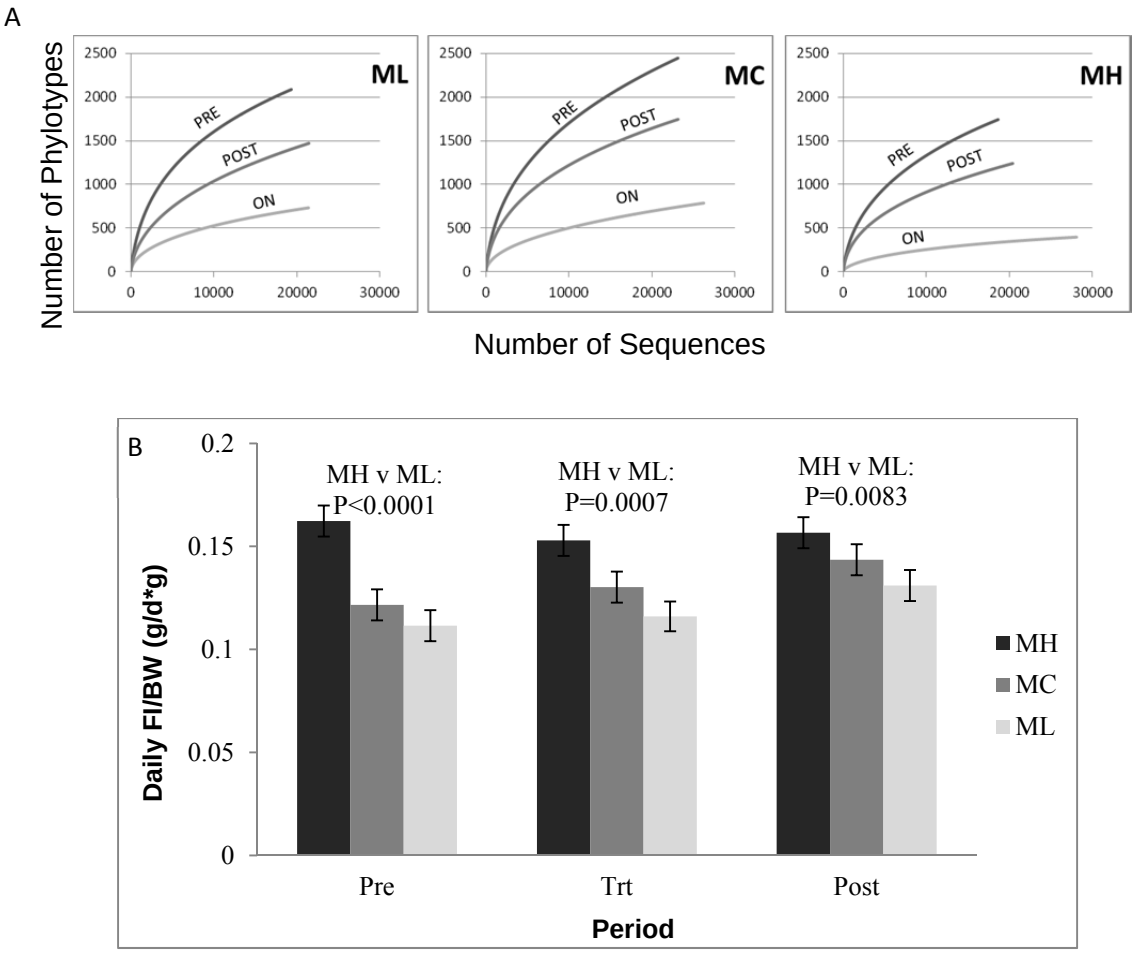

Figure 2 (A,B): Effects of antimicrobials on feed intake in mice selected for heat loss characteristics. Large doses of antimicrobials were administered to selection lines to perturb the line specific microbiotas. (A) Fecal pellets from four animals from each line were selected for 16S rRNA-based pyrosequencing over the time course. Rarefaction curves based on complete linkage clustering of OTUs measured the effects of antibiotics on overall diversity of microbial communities throughout the treatment. OTUs were chosen at a cutoff of 97% sequence identity. (B) The effect of antibiotics on feed intake was monitored in 167 animals. Graphs depict mean daily FI/BW and bars show standard errors. The pre-treatment difference in feed intake of 41.8% (P <0.0001) between lines was reduced to 28.3% (P=0.0007) during antibiotic treatment.

	Period		
Selection Criteria	Pre Treatment	Treatment	Post Treatment
MH	0.1623	0.1529	0.1566
MC	0.1216	0.1302	0.1435
ML	0.1115	0.116	0.131

Table 5: Feed intake per BW of Animals under antibiotic treatment. Average feed intake per body weight (g/d*g) for the line across the antibiotic treatment period (Pre, Trt, Post).

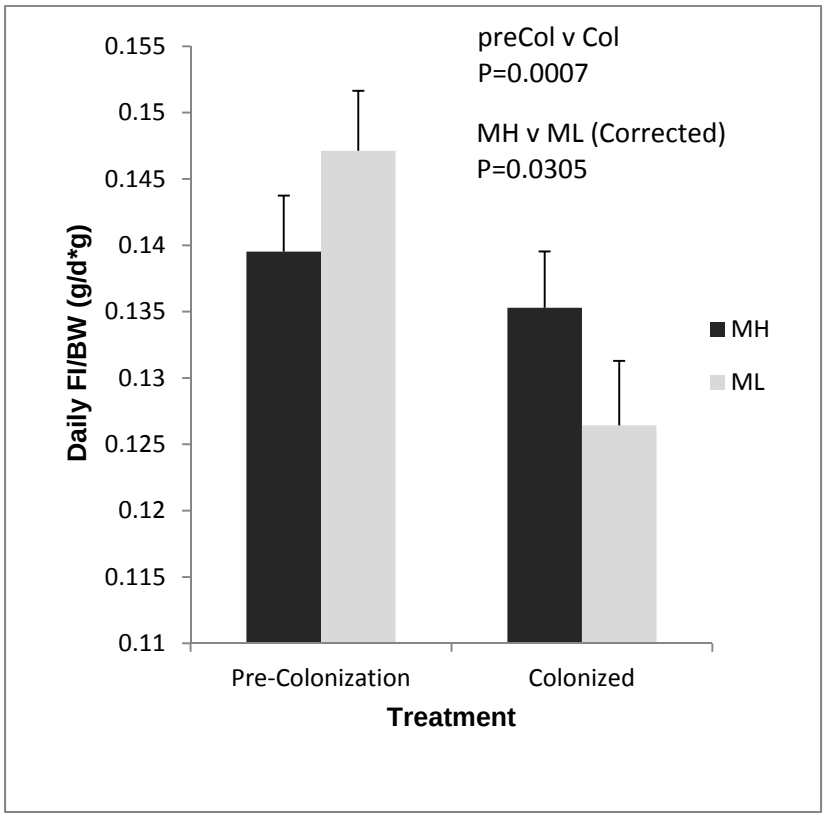

Figure 3: Transfer of feed intake traits into Germ-Free recipients using fecal material from donor MH and ML selection lines. Recipient germ-free Swiss-Webster mice were maintained in two separate germ-free isolators. Graphs depict mean daily FI/BW and bars show standard errors. PreFI/BW for either isolator (P > 0.25) whereas PostFI/BW was statistically highly significant (P=0.0305); when adjusted for pre feed intake to body weight (PreFI/BW). Mice colonized with microbiota from MH consumed 0.136g feed/gBW/day whereas mice colonized by ML microbiota yielded only 0.126 g feed/g BW/d, se = 0.003)

Period	MH	ML
Pre-Colonization	0.139528	0.147118
Colonized	0.132869	0.130002

Table 6: Feed intake per BW of Animals of gnotobiotic animals before and after line specific microbiome colonization. Average feed intake per body weight (g/d*g) was measured for the lines across the gnotobiotic study. Pre-colonization levels reflect isolator-specific variation. Colonized is average feed intake per body weight after Swiss-Webster mice were introduced to the line specific microbiota for 4 weeks.

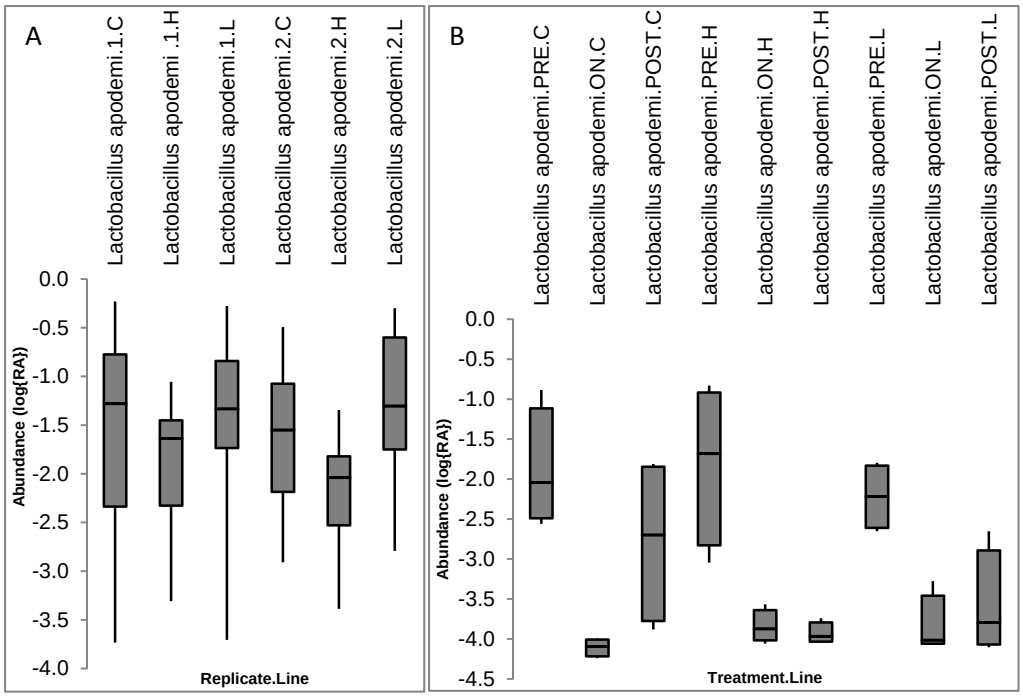

Figure 4 (A,B): Box and Whisker plots of *L. apodemi* levels in untreated and antimicrobial treated lines.. Box and whisker plots depict 75% of values (box) and range (whisker) with bars as the mean. (A) Untreated animals: Replicates (1,2) showed *L. apodemi* to be significantly higher (P<.05) in the ML (1.L, 2.L) line than the MH(1.H, 2.H) with the intermediate MC (1.C, 2.C) line. (B) Antibiotic treated animals: In all lines(.C, .H, .L) antibiotic perturbations (.PRE→.ON) nearly eliminated *L. apodemi* corresponding to an intermediate feed intake characteristic. With the large antimicrobial dose recovery of *L. apodemi* appeared strained (.ON→.POST) and feed intake did not consistently return to pretreatment levels after treatment.

Selective breeding for feed intake characteristics in mice results in unique gut microbial communities that contribute to feed intake phenotypes.

Ryan Legge[1,3], Merlyn K. Nielsen[2], Adrienne Bhatnagar[2], Rhonda Griess[2], Jens Walter[1], Daniel A. Peterson[1,4], Jaehyoung Kim[1,3], and Andrew K. Benson[1,3,*]

[1]Departments of Food Science and Technology, [2]Animal Science, and [3]Core for Applied Genomics and Ecology, University of Nebraska, Lincoln, NE 68583-0919; [4]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205-2109;

3. USE OF PYROSEQUENCING TO IDENTIFY NEW INDICATORS OF FECAL CONTAMINATION AND TEMPERATURE ABUSE IN LEAFY GREENS.

3.1. ABSTRACT

Microbiological testing is a primary strategy used to assess risk that foods and beverages may be contaminated. Though quite sensitive, current microbiological methods rely on cultivation of specific organisms or groups of organisms in order to elevate the population and/or to allow visualization and enumeration. A major approach to risk estimation has been reliance on "indicator" organisms that are believed to be associated with risk of pathogen contamination. Reliance on these cultivation-dependent methods poses many limitations, one of the most significant being the very limited number of taxa that can be cultivated and enumerated. Even methods which enumerate the entire microbial load provide no information about composition, and show only limited correlation between absolute numbers and food safety/quality characteristics. To circumvent these problems, this project utilizes cultivation-independent, community DNA sequence-based methods as a means for safety and quality assessment of foods. Using a model system of fresh spinach leaves, DNA sequence-based strategies were used to test several criteria that would need to be met in order to use community analysis as an alternative to standard microbial screening. Results indicate that pyrosequencing of 16S rRNA tag sequences amplified from spinach samples can readily differentiate the microbiota of spinach from the microbiota of bovine feces, a primary source of contamination. Data shows compelling evidence that the microbiota of spinach is predictable; it is

distinct from fecal microbiota; and predictable correlated changes occur when spinach is contaminated artificially in the laboratory. Thus, cultivation-independent, DNA sequence-based approaches is an alternative to culture-based microbiological testing.

3.2. INTRODUCTION

Food borne illnesses comprise nearly 47.8 million cases each year in the United States. This translates into the striking statistic that nearly 1 in 6 Americans will have an incident of food borne illness each year. Even though most cases are mild and often cause symptoms indistinguishable from other causes, the CDC estimates that there are at least 127,000 hospitalizations and 3,030 deaths related to food borne diseases each year. (215) From 2006-2008 3,401 foodborne outbreaks have been reported (216-218), with 42 large, multistate outbreaks occurring over the last 5 years (215). Though system-oriented approaches such as Hazard Analysis and Critical Control Points (HACCP) have been implemented widely, culture-based microbiological methods are still used to mitigate risk of contamination. Because of the variety of enteric pathogenic species and the fact that they are often present at low levels, indicator tests such as total coliform count, *Escherichia coli*, or total aerobic plate count method (219) have been used to assess relative risk that a sample is contaminated with a pathogenic species or spoiled.

Historically, indicator organisms (*E. coli* and coliforms) were thought to comprise the dominating members of the intestinal microbiome because they are able to be cultured, analyzed, and enumerated (25). However, the vast majorities of microbial species are unable to be cultured or grow under strictly anaerobic conditions and are

therefore difficult to culture (220-223). With the development of culture-independent approaches and improved methods for culturing anaerobes, fecal coliforms were shown to make up only a very small portion of the overall intestinal microbial mass. All total facultative aerobes, including *E. coli*, make up less than 0.001% of the microbiota in the strictly anaerobic mammalian colon (32, 111). Furthermore, dominate members of the intestinal ecosystem outnumbered currently used fecal coliforms by many orders of magnitude.

To further complicate their status as risk indicators of fecal contamination, many coliform species occur naturally in environments outside the gastrointestinal tract; some are even known to be part of the epiphytic microbiota of plants (113). Coliforms such as *Klebsiella*, *Citrobacter*, *Enterobacter* (among others) are also found in soils and can be observed in degrading plant material (219, 224, 225). Others such as *Bacillus*, *Sphingomonas*, and *Aeromonas* can cause false positive results in coliform testing although they are not considered part of the coliform group (226). Given the widespread environments inhabited by coliforms, the significance of their presence in with raw commodities or minimally processed foods is debated. "Generic *E. coli*" does have some merit as an indicator in specific situations such as estimating risk that bovine carcasses are contaminated with fecal material post slaughter. The USDA Food Safety and Inspection Service (FSIS) therefore have included this in mandated HACCP programs for meat processors (USDA Food Safety and Inspection Service (FSIS)). Nonetheless, because *E. coli* comprises only a fraction of the fecal microbiota, one questions the sensitivity of risk assessment approaches based on this

indicator. Moreover, when generic *E. coli* testing has been used retrospectively to test pathogen-contaminated foods, poor correlation is often observed (227).

Compounding the food safety issues is the fact that economic drivers and market demand is causing food production companies to increase automation, rate of production, and consumers are demanding ready to eat or minimally processed products. In addition importation of raw commodities such as fruits and vegetables, has increased. Because of this, increasing responsibility lies on the producers and production companies, making them eager to assess new technologies to limit risk (228). Studies to date have focused on ways to rapidly detect specific pathogens in food products focusing on the use of molecular methods to confirm presence/absences in enrichment cultures. Very little emphasis, however, has been placed on bettering approaches of indicator testing (113).

Given this opportunity and the availability of methods for detailed analysis of entire microbial communities, we have focused on evaluating community analysis as an approach to augment or even replace traditional indicator testing. With outbreaks through food borne pathogens being linked to fecal-oral transmission (25, 229, 230), fecal indicator testing remains an important tool for risk assessment. Using a microbial community-based approach with non-culture-based methods, species not readily culturable by standard methods but which are present at orders of magnitude higher than *E. coli* can be used to asses risk of fecal contamination.

Micro- and macro-organisms are habitually associated andiInteractions between them shape contrasting environments in different host-microbial communities (178). Furthermore naturally occurring populations of bacteria and their host, including

interactions between foodborne pathogens, contribute to entophytic and epiphytic colonization in plants (114). Currently, 454-based pyrosequencing has become an established methodology for examining the composition and abundance of diverse microbial environments. The methodology can measure compositional features of complex microbial communities in detail, providing an environmental snapshot for microbial ecosystems previously unattainable. DNA based sequencing approaches make it possible to explore microbial communities which were under-sampled or previously unknown. (16, 40, 41, 50, 108, 109). Broader dynamic ranges achieved from this method (111) increases our ability to detect if a community includes a given species characteristic of other environments such as feces, sewage or soil and in turn could be used to assess quality. In this aspect, sequencing technology could be practical in food quality settings to test for microbial signatures that should not be present in wholesome foods. Pyrosequencing of 16S rRNA amplicons allows for detailed characterization of the microbial community from total DNA extracted from a food source. The community profile results from massively parallel DNA sequencing of these 16S rRNA amplicons so that each sequence read from a sample represents an individual microbe. Bioinformatic and statistical analysis then allow for robust hypothesis testing and conclusions about a single taxa or the entire community.

Recent outbreaks of food borne illness have been attributed to leafy greens and other fresh fruits and vegetables. It was not until the 1990's that these foods became a safety concern, an issue that underscore the need for increased surveillance and methods that can detect pathogens in unsuspecting food, unusual or new pathogens, or simply assess risk of contamination. Spinach is an ideal model as it is a minimally

processed, ready-to-eat, raw commodity and it has been on the forefront of food safety since the last major outbreak in 2006 (215).The approach will test the hypothesis that spinach flora has unique microbial communities that differentiate it from the fecal microbiota in bovine populations. Secondly, we will test the hypothesis pyrosequencing of 16S rRNA amplicons can detect contaminated or abused samples.

3.3.MATERIALS AND METHODS

3.3.1. Experimental samples, survey protocols and sample collection.

3.3.1.1. *Spinach samples:* Freshly harvested spinach leaves were obtained from three different locations in Salinas Valley, CA and one location in Yuma, AZ. A total of 111 leaf samples were obtained, 72 leaf samples from sites G,R,S in Salinas Valley, CA and 39 from site Y in Yuma, AZ. These samples were assessed cover variations in independent variables (such as cultivars, growing season, climate, processing, etc.). Salinas Valley and Yuma represent the two primary regions of leafy green spinach (>90%) in the United States during the summer (Salinas Valley) and winter (Yuma). Having direct access to plants from Salinas Valley, CA and Yuma, AZ gave an unprecedented look into food grade raw commodities these regions are principle in supplying the leafy greens to the United States. Samples from these regions were shipped on dry ice, levas were washed to remove the

phylloplane bacterial populations and DNA was extracted from this wash immediately upon arrival.

3.3.1.2. *Bovine fecal samples:* To develop represent bovine fecal microbiota profiles, 145 bovine feces from the USDA-MARC herd Nebraska were collected by rectal grab and stored at -80 until use for DNA extraction.

3.3.1.3. *Corn samples:* Direct sampling of corn samples was completed to determine if variation within species was greater than between maze and spinach. The corn samples were obtained from 77 different lines of the NAM (nested association mapping) collection (231-233), a genetic resource population.

3.3.1.4. *Sampling for temperature abuse and cross contamination of feces on spinach:* Cut spinach leaves from 50 randomized G,R,S samples in Salinas Valley, CA were pooled to achieve a uniform composite. Samples of ~20 leaves were chosen from the composite and subjected to three independent bovine fecal contaminations from the diverse bovine fecal samples from the USDA-MARC herd. Fecal contamination was presented though a slurry by making dilutions of 1:100, 1:1,000 or 1:10,000 by diluting feces weight to volume in 0.1% peptone solution. Contamination with each of these dilutions was carried out in two different contamination ratios; 1:1 and 1:3 contaminated to uncontaminated leaves. Portions of each dilution and ratio

treatment were sampled over 4 days to monitor effects of contamination over time. All samples were stored at 4ºC during the experiment. DNA was extracted immediately after sampling at each time point over the 4 day study. Positive controls were made by directly inoculating 2g of fecal material into 20 leaves of spinach, negative controls for spinach used 20 leaves from the composite. The fecal samples were also sampled directly. A portion of the spinach composite was also used to test for effect of temperature abuse on the microbiota. For this experiment, samples of 20 leaves were incubated at roughly 25ºC (Room Temperature) and 37 ºC for the 4 day period and sampled each day over the time course. A blank negative control was used containing only the peptone solution. This blank negative control represented 0.00196% of the total reads signifying a low background for the experiment prior to the sequencing run.

3.3.2. Whole DNA extraction.

Microbial communities are assessed by washing the epiphytic surface of cut spinach in 50mL 0.1% peptone solution for 30 minutes. After washing, the buffer was removed and the bacteria were recovered from the buffer by centrifugation. Total DNA extraction is then completed in parallel on the BioSprint 96 (Qiagen) using the BioSprint 96 One-For-All Vet Kit. The procedure is followed per manufacturer's recommendations except that an additional step of physical lysis is introduced by bead beating with glass beads and the TissueLyserII (Qiagen).

3.3.3. Pyrosequencing.

All 111 spinach, 145 bovine, and 77 corn samples were subjected to 454-based amplicon pyrosequencing. The sequencing was conducted on PCR products amplified from the V3-V1 region of the 16S rRNA genes. The region was amplified from the total DNA extract using modified versions of the A518 and B8 PCR primers. The modified primers contain the Roche "A" or "B" sequence adapter upstream of the A518 and B8 sequence. A unique 8-bp sequence was added to 96 different "A" primers so that multiple "A" primers could be used to amplify individual samples and resulting samples could be sequenced in parallel. The V3-V1 region of the 16S rRNA gene was amplified using bar-coded fusion primers with the Roche-454 A or B titanium sequencing adapters (*in italics*), followed by a unique 8-base barcode sequence (B) and finally the 5' ends of primer A-518RM (5' - *CCATCTCATCCCTGCGTGTCTCCGACTCAG*BBBBBBBB ATTACCGCGGCTGCTGG-3') and of primer B-8F (5'- *CCTATCCCCTGTGTGCCTTGGCAGTCTCAG*AGAGTTTGATCMTGGCTC AG - 3′). PCR conditions (TaKaRa Ex Taq) and pyrosequencing runs (454 Roche) follow the manufacturers' recommendations. All PCR reactions were quality-controlled for amplicon saturation by gel electrophoresis; ban intensity was quantified against standards using GeneTools (Syngene). For each region of a two-region picotiter plate, amplicon reactions were pooled in equal amounts based on the GeneTools output to achieve a read distribution of

~10,000 sequences/sample and the resulting sample was gel-purified (16). Recovered products were quantified using picogreen ds DNA broad range assay (Invitrogen Q32850) by a Qubit fluorometer (Invitrogen Q32887), spectrophotometer (nano-drop ND-1000) and bioanalyzer (Agilent 2100) and sequenced using Roche-454 GS FLX/Titanium chemistry. Raw read output was filtered by length and quality procedures and binned by sample specific barcodes. For each region of a two-region picotiter plate, amplicons from up to 48 reactions were pooled in equal amounts.

3.3.4. Raw data filtering and binning.

Raw read data from the 454 pyrosequencing runs were processed through a quality filter removing sequences that fail to meet the following criteria:

- A complete forward primer and barcode present
- No more than 2 "N" characters (where N is equivalent to an interrupted and resumed signals from sequential flows)
- Length greater than or equal to 200nt but not longer than 500nt
- An average quality score above 20

After filtering, sequences were binned to sample-specific barcodes. Each read is trimmed to remove 3' adapter, primer sequences, and barcode. The corresponding FASTA and QUAL files were updated to remove quality scores from reads not passing quality filters. The files are associated with sample information in a hierarchical manner in MySQL tables. The processed data

and the MySQL database tables are stored on a database server http://cage.unl.edu, allowing data to be made public after publication.

3.3.5. Taxonomic analysis and statistical method.

A CLASSIFIER+CD-HIT approach was subsequently used for determining taxonomic classification. Sequences were first parsed through the MULTI-CLASSIFIER (131) to separate those sequences that meet a genes-level threshold criterion. The algorithm assigns taxonomic status to each sequences read based on a covariance model developed from a training set of 16S rRNA sequences. Reads are classified down to the genus level at a threshold of >0.8 were further assigned species or OTU status using the best-BLAST hit to search a highly curated set of reference sequences from RDP CLASSIFIER (131) and the SILVA (198) databases. Sequences not meeting the BLAST criterion were assigned as an OTU with a genus level taxonomy. Sequences unable to be CLASSIFED at any taxonomic level were concatenated and binned through an OTU picking approach CD-HIT-EST (143). This algorithm groups related sequences on the basis of k-mer similarity by a cutoff threshold into OTUs for downstream analysis. A .97 sequence identity threshold and a .90 minimum coverage threshold were used to model 'species'- level phylotypes. As with classified sequences, the OTUs from CD-HIT were also used to query the curated 16S rRNA database to assign best-hit taxonomic status. OTUs or taxonomic bins were excluded if the bin failed to have on average 20 sequences per animal for taxonomic bins

above this minimum, any animals having no counts were assigned 0.5 read abundance to preclude issues with log transformation.

After assigning taxonomic status through the CLASSIFIER/OTU pipeline, absolute proportions of communities meeting this criterion were conducted by:

$$absolute\ proportion = \frac{Number\ of\ reads\ in\ a\ taxon}{total\ number\ of\ reads\ in\ a\ sample}$$

Communities were excluded if the overall environmental group (corn, bovine, or spinach) or (Contaminated, Control, or Heat) per taxonomic assignment is 0.1% of the average relative proportion across the group defining a minimum abundance threshold for detection. Taxa below this threshold were discarded if the threshold was not met for at least one environment tested. In this regard taxa may be below the threshold or even zero as long as a second environment was above the 0.1% baseline. Significance testing was conducted between groups by one way ANOVA treating environmental group (spinach leaves, corn leaves, or bovine fecal isolates) as the main effect. When spinach leaves were treated with bovine fecal isolates or temperature abused all polluted samples were treated as contaminated or temperature abused, respectively, regardless of individual treatments applied to the leaves. Scheffe testing was performed within taxa to test specific contributions between environmental groups. All significant members of the microbial communities across environmental groups were determined significant ANOVA at $p<0.01$. PCA

analysis was performed across all samples on species level taxonomic assignments crossing the minimum threshold of 0.1%.

3.4.RESULTS

3.4.1. Technical Repeats of Spinach Composites

Because the epiphytic microbiota has highly complex communities we first tested repeatability of sequencing to determine the threshold for sample error. Technical repeats of four identical samples spinach composite showed a relative abundance of 0.1% was necessary for any given taxonomic group to have a high probability of correlation between replicates (Figure 1). This threshold was applied across each group within the entire population. While the taxa meeting this group represented only a small proportion of the species (<2%) detected they accounted for >85% of the sequence reads assigned to a species level by the CLASSIFIER+CD-HIT method. This minimum abundance threshold provided the baseline that taxa must be present on spinach or when leaves were contaminated/abused to be useful for identifying signatures within spinach microbiota or across the microbiota of abused leaves or contamination events.

3.4.2. Core microbial compositional differences of spinach and corn leaves to bovine feces.

To serve as a measure of contamination, the microbial community from the food and fecal matrices must have compositional features that are unique and differential. The minimum abundance threshold selected for taxa across 13 phyla, 20 classes, 35 orders, 77 families, and 152 genera with our CLASSIFIER+CD-HIT pipeline identified 424 usable species (meting the minimum abundance threshold) across the corn, bovine, and spinach environments. Our pyrosequencing results showed that even at the phylum level, differences between fecal and plant (corn and spinach) microbiotas were substantial. Discriminate power between sample types using PCA increased at lower taxonomic levels (Figure 2). In general, plant members contain large proportions of Proteobacteria, Actinobacteria and Cyanobacteria (which likely arise from chloroplast DNA contaminating the DNA sampled). Fecal samples showed the expected large proportions of organism belonging to the phyla Bacteroides and Firmicutes (Figure 3), which are generally found with low abundances within the spinach flora. At the family level, spinach was dominated by Bacillaceae, Enterobacteriaceae, and Pseudomonadaceae these make up, on average, 33.5% of the sequencing reads across the spinach phylloplane. The spinach and corn phylloplanes were differentiated from one another, but share the dominant taxa Enerobacteriacea and Psudomonadaceae, in corn these taxa made up 17.3% of the average sequences. Rhizobiaceae, Flavobacteriaceae, Comamondaceae, and Xanthomonadaceae also dominated

the corn environment at each >5% of the total average sequencing reads combined these taxa make less than 1.6% of the average sequence reads in spinach. The above six taxa dominating corn phylloplane reads encompass and 55.1% of the sequencing reads of corn. Furthermore when taken into account that chloroplast reads are picked up in these two environments, these dominating taxa make up 55% of the non-chloroplast reads in spinach and nearly 66% of the non-chloroplast reads in corn.

Both of these plant microbiotas were readily distinguished from bovine feces which, on average, contained high numbers (>4.9% of the total average reads) of Lachnospiraceae, Provotellaceae, Ruminoccaceae, and Succinovibrionaceae that are barely present on the spinach phylloplane; together these taxa make up less than 0.08% of the spinach phylloplane. Collectively these made up 61.7% of the average total bovine sequencing reads. Statistical analysis showed clear association with the host as a total 11 phyla, 18 classes, 31 orders, 71 families, 133 genera, and 354 species (including all the above mentioned dominating taxa members) were significant by ANOVA at $p<0.01$. Thus, the first criterion seems to be fulfilled; the plant phylloplane microbiota is clearly distinct from the fecal microbiota.

3.4.3. Contamination and abuse of spinach.

To determine if pyrosequencing could detect changes in the microbiota from fecal contamination or temperatures abuse, a pool of spinach leaves was

prepared and contaminated at various levels with a fecal slurry (see materials and methods) or was subjected to simulated temperature abuse (storage at room temperature and 37ºC). Using a minimum abundance threshold of 0.1% taxa, 307 usable species were established, representing 10 phyla, 16 classes, 26 orders, 60 families, and 104 genera with our CLASSIFIER+CD-HIT approach. To identify taxa that were true indicators of contamination or abuse, contaminated and temperature abused leaves were tested at all time points by ANOVA; 4 phyla, 7 classes, 14 orders, 23 families, and 38 genera and 112 species were found significant ($P<0.01$).

3.4.3.1. *Detection of bovine feces on spinach:* Given the uniqueness of bovine fecal microbiota we tested both ecological based approaches and statistical approaches for their ability to discern contaminated samples. For our ecological approach, we used diversity (species richness and evenness) estimates to determine if diversity changes could be used to signal contamination. Diversity was estimated, pyrosequencing data by rarefaction analysis (131) (Figure 4).

Control spinach samples in general contained less phylotypes per sequence read than highly contaminated leaves. This might be expected as these two environments have unique microbitas and mixing them would theoretically increase the overall diversity. However, samples with dillute fecal material and low ratios of contaminated to uncontaminated leaves less diversity was observered and the samples are not left shifted as with high contamination. There was also variation depending on which fecal samples

was used for the contamination. Fecal samples 1 and 3 show nearly all contaminated samples as left shifted but Fecal sample 2 shows only a subset as it had lower levels of contamination. Regardless, any sample with a diversity greater than 2000 phylotypes at a depth of 5000 sequences was clearly contaminated. In this regard diversity alone could potentially be useful in detecting contamination, as samples above this threshold are more likely to have encountered contamination.

As a second approach we again used pattern discovery methods, such as PCA, to determine if microbiota from contaminated leaves was differential from control leaves. PCA analysis on 442 species level OTUs meeting the minimum abundance threshold separate contaminated samples (Figure 5). The discriminate power increased at lowest taxonomic levels with PCA of species/OTU data showing grouping of sample by type. Variation in the contamination patter was primarily due to storage at 4ºC for the time course as well as different dilutions of feces applied to leaves.

Significance testing by ANOVA also showed clear differentiation in the microbiota of contaminated, and control spinach leaves. The distinction between contaminated and control leaves was primarily within the phylum Firmicutes, where contaminated leaves had on average 49.2% of their sequencing reads dominated by this phyla compared to >7.5% in the spinach control. Nearly all of these reads (64% or 31.4% of all contaminated sequence reads) were comprised from the five families of Lachnospiraceae, Clostridiaceae, Peptostreptococcaceae, Ruminococcaceae, and

Erysipelotrichaceae (Figure 7) each having on average >4.9% of the sequencing reads for the contaminated spinach leaves. In contrast control leaves were primarily dominated by the Pseudomonadaceae comprising nearly 31% of all sequencing reads for control spinach. These taxa dominate the fecal microbiota and are several orders of magnitude more abundant that *E. coli* (generic) which wasn't even detected in our samples. Given their abundance, specifically for fecal environments, and stability at 4ºC, we believe these taxa could serve as a new generation of true fecal indicators. qPCR assays specific for these taxa are likely to be much more sensitive and reliable predictors of risk.

3.4.3.2. *Temperature abuse of raw spinach leaves:*

When looking at temperature abuse specific taxa among the spinach flora expanded compared to controls. Overall quantitative species composition, PCA analysis was again able to separate abused and control samples. (Figure 6).

Taxa responeded to temperature abuse were defined as having >1% of the average sequencing reads for temperature abused samples and >3.5 fold difference in relative abundance in relation to the spinach controls. The families of Sphingobacteriaceae (4.6% of the average sequencing reads, 9.12 fold difference average of temperature abused leaves over control leaves), Alcaligenaceae (1.2% of the reads, 15.2 fold difference), Comamonadaceae (1.2% of the reads, 6.6 fold difference), Flavobacteriaceae (6.9% of the reads,

19.6 fold difference), and Bacillaceae (3.4% of the reads, 3.8 fold difference) were also greatly expanded when leaves were temperature abused (Figure 7). The relative ratios of microbes from these families could therefore serve as markers to define the presence of improper storage as these taxa continued to expand over the 4 day sampling period.

Results from these two studies showed clear distinguishing markers between wholesome and contaminated/temperature abused samples.

3.5.DISSCUSSION

With next generation sequencing platforms such as the Roche 454 Flx+, the Illumina HiSeq, and Ion Torrent PGM rapidly evolving, sequencing will likely be used in the near future on a regular basis in diagnostic testing. It is now possible to both qualitatively and quantitatively evaluate microbial environments that are associated with food and food product testing. Furthermore, massively parallel sequencing produces community snapshots that are not possible by using traditional culture-based methods. With the broad dynamic range produced by pyrosequencing, new candidates that accurately predict fecal contamination or temperature abuse can be identified across a wide variety of food products. Ideally, defining candidates could lead the forefront of new rapid method testing in predicting wholesome verses contaminated foods. This approach can help pinpoint more accurate indicator taxa. Moreover, these indicators can be established for risk as well as favorable characteristics such as shelf-life or even organoleptic properties.

Findings linking community structure to the natural ecology of plants could help understand the processes behind colonization of fresh produce and conditions that may favor colonization or persistence of pathogens. It is possible that the natural plant microflora could reduce, inactivate or inhibit enteric pathogens in raw commodities. Information gained will be advantageous in developing agricultural practices for pre-harvest and post-harvest safety of fresh fruits and vegetables (114).

The sequencing technologies can allow for visualization of all the changes that take place when the genetics, nutrition or environment of food plants, animals, or production facilities are altered (234). Identifying pathogens directly is difficult as many pathogens are rare and detection is limited to highly technical, low level clinically based PCR assays. To further complicate things, only a small group of genes differentiate virulent strains from non-virulent which would be grueling to isolate and monitor over large samples (111). Moreover, as illustrated by the newly emerged *E. coli* O104:H4 outbreak in Europe. New strains can emerge quickly that are not detected by current testing methods. Using a true metagenome study, along with assembly of the sequences into long genomic segments of the organisms, one could potentially differentiate known pathogens and strains having new and potentially virulent gene combinations (235). Such an approach would also allow one to study the ability of microbes to adapt to and survive various processing conditions (114). Second generation of DNA-sequencing platforms will become commercially available in near future; these instruments will provide a new depth of sequencing at a lower cost than previous generations. These new machines will provide the sequencing power to limit the use of traditional 16S rRNA amplicon sequencing and

instead allow for true metagenomic studies of environmental isolates. Metagenomic studies will be able to provide important insights into the evolution, biology, and ecological fitness of microbial communities and the genetic functional properties within them. Genomic tools could be expanded directly and quickly monitor for food pathogens, or virulence and spoilage genes. Furthermore, technology will allow for sequence based studies assessing bacterial behavior across the entire food production process, from the raw material to the finished consumer products. This will provide precise data on how communities respond to production stress and where potential hazardous microbes are entering the food stream. (113) In turn precise mechanism to control food safety and quality may be determined at the microbial level.

However, currently to adjunct or replace indicator testing models with community profiling methods would mark a sizeable leap in food assessment. However, the direct use of pyrosequencing as the sole method for assessing risk is highly technical and time consuming, requiring an experimentally driven foundation in laboratory settings. Even without these hurdles, the technology is still far too expensive to replace indicator testing. However, the pace at which the sequencing technologies are evolving portends eventual closure of the cost hurdle. Because bioinformatics will be the barrier, the time is right to begin development of bioinformatics tools that can capitalize on sequencing platforms.

In practice it direct assessment could be completed in industry settings, however it is unlikely that enough users could have the requisite expertise in the food industry and the time and computational constraints would make it difficult to warrant widespread use. Therefore, to implement risk management techniques, identified

candidates from sequencing relating to a high risk of contamination/abuse will need to be analyzed through other methods. Quantitative real-time PCR is an attractive application for this as it allows for thousands of different reactions to be completed simultaneously within a few hours. If high risk microbial candidates can be identified and established, future work would focus on implementing alternative methods that could be utilized in the food industry. Our data identified members the Families of Lachnospiraceae, Clostridiaceae, Peptostreptococcaceae, Ruminococcaceae, Erysipelotrichaceae Sphingobacteriaceae, Flavobacteriaceae, Alcaligenaceae, Comamonadaceae and Bacillaceae. Many of these families identified in this study as indicative of contamination have already been analyzed and quantified by qPCR methods in laboratory settings (236-238), this leap into the indicator testing arena would therefore be effortless. Furthermore, significant members of our lab contamination of spinach leaves included Lachnospiraceae, Clostridiaceae, and Ruminococcaceae, members that have been shown to be "ubiquitous bacteria detected from cattle feces" (239). Furthermore, Erysipelotrichaceae has been found to be a high abundance member of dairy cows (240) and Peptostreptococcaceae has been associated with yearly fluctuations in bovine feces (241). Family members of Flavobacteriacea (242) and Bacillaceae (243) have also been documented to follow plant destruction and spoilage, giving credibility to our findings.

The used pattern-finding approaches such as PCA to differentiate samples and help identify taxa contributing contamination and abuse. The advantage is that the entire set of taxa was considered as opposed to treatment of taxa as discreet units in

ANOVA. Use of PCA-type methods may therefore help identify natural trends in the data and be less subject to multiple testing errors associated with ANOVA.

Organisms found in the top percentile of PCA case wise scores were extrapolated. These organisms played a major role in shaping the overall PCA pattern finding. Top scores for contamination organisms such as species of *Faecalibacterium prausnitzii*, *Roseburia intestinalis*, and OTUs of genera from *Turcibacter*, *Oscillibacter* have been characterized as members of the gut of feed animals (244-246). Spoliage organisms such as species from the *Bacillus*, *Pseudomonas* and *Erwinia* genera (243, 247, 248) were also found as being in the top percentile of case wise PCA scores for temperature abuse.

Initial findings signify that 454 sequencing utilizing community profiling methods could be implemented as a novel method of assessing new indicator organisms in raw commodities. With outcomes clearly showing distinction between wholesome and contaminated products community profiling could lead the forefront for a host of novel molecular methods in eliminating high risk foods from ever reaching consumer markets.

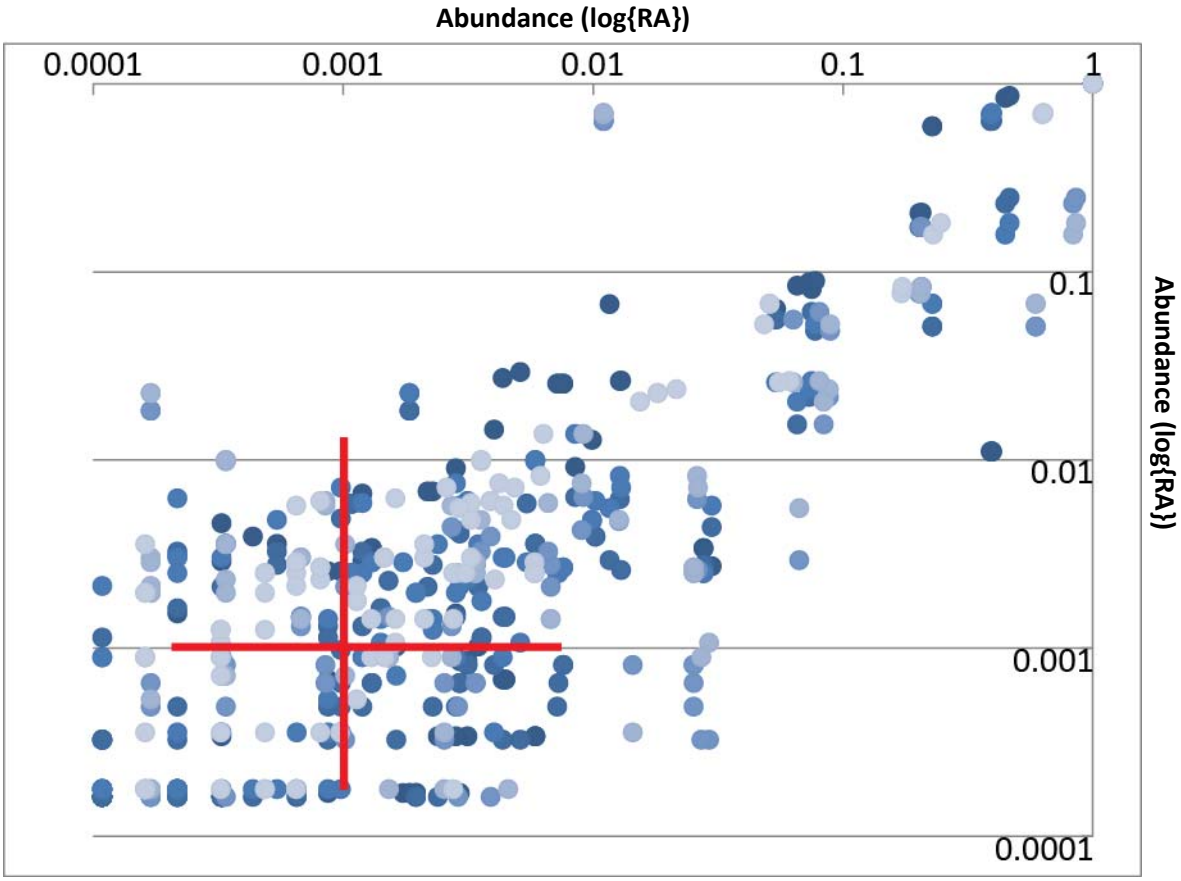

Figure 1 - Pairwise combinations of data from four spinach biological replicates. Processed and filtered sequences from each barcode–sample combinations were then assigned taxonomy by CLASSIFIER. Relative abundance (RA) for sequence counts in each taxonomic bin was plotted for all pairwise combinations of the replicates. Axes were log10-transformed values for total sequence reads of each taxon. The red crosshairs indicate the 0.1%-average read threshold. Above this number, correlation reaches >0.741 significant at p <.001: below this number, correlation dissipates rapidly.

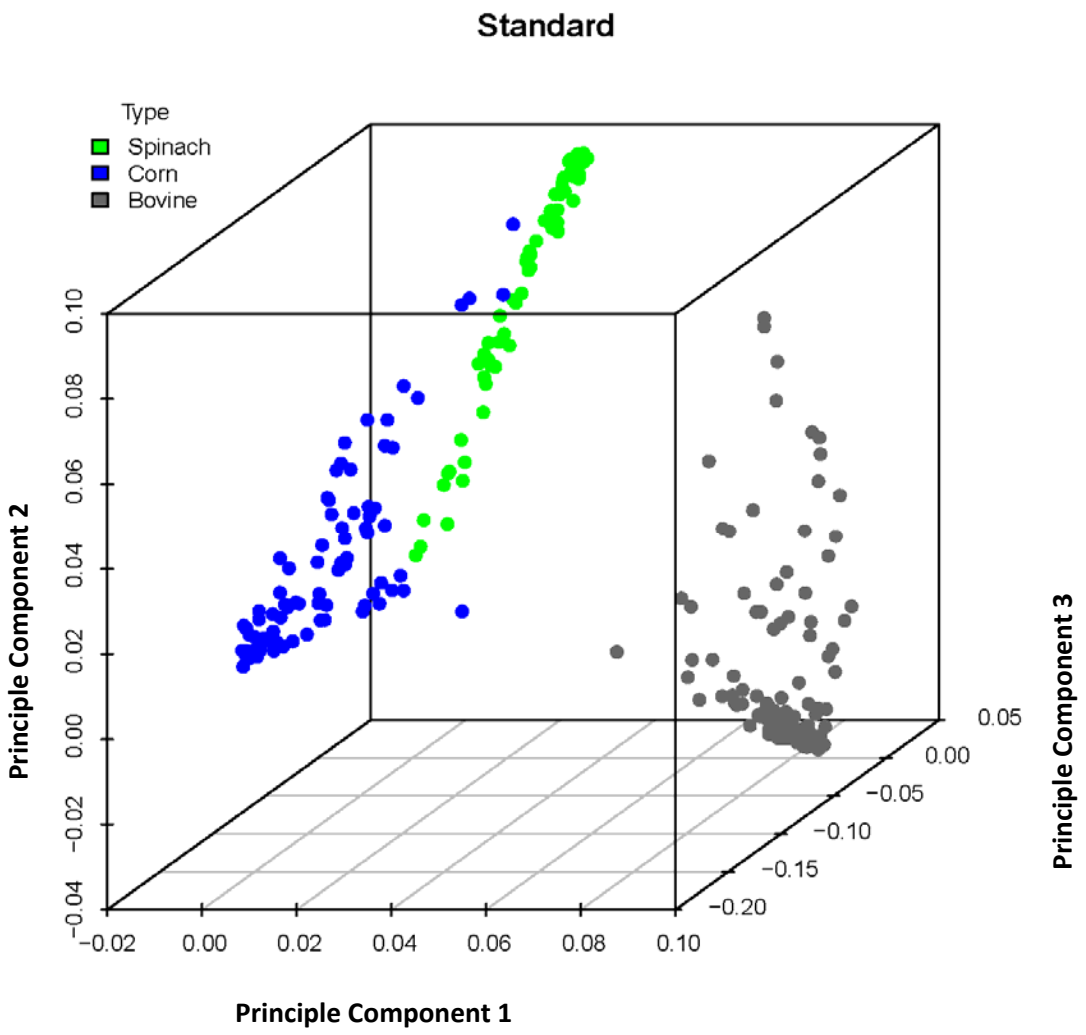

Figure 2 - PCA grouping of species level taxa between bovine feces, corn leaves, and spinach leaves. 442 species were used to conduct PCA analysis across all samples for the three environments above the minimum abundance threshold of 0.1% (Corn Leaves, Spinach Leaves, and Bovine Feces). The additive effects of species explaining the variation of these environments is primarily in the first three principal components (1 – 32.86%, 2 – 26.54%, 3 – 9.25%).

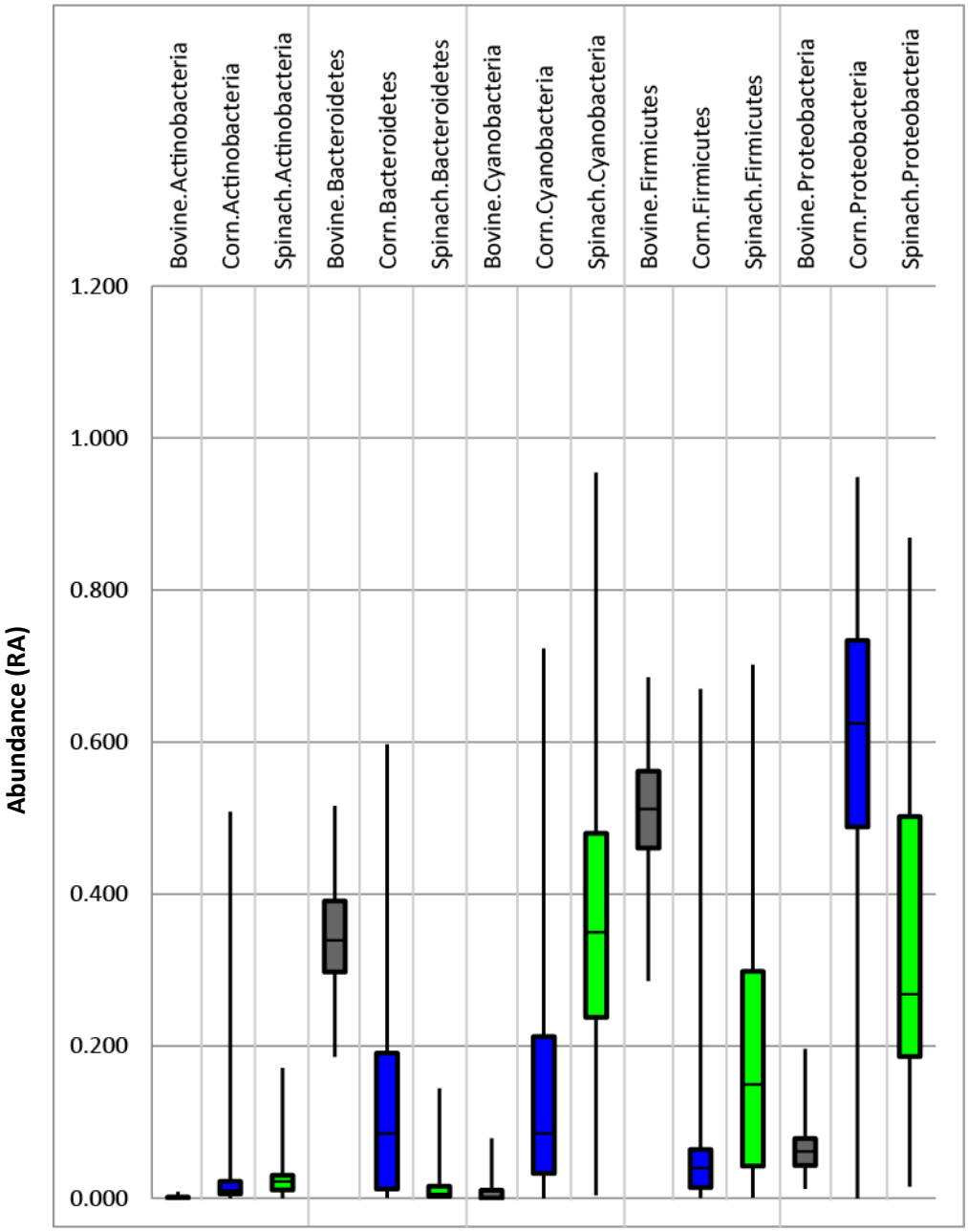

Figure 3 – Box and whisker plots of bovine feces, corn leaves, and spinach leaves. Taxonomic reads assigned to the phylum level were plotted where box and whisker plots show: the maximum, 75 percentile, mean, 25 percentile, and minimum relative abundance (RA) across each environment. Bovine Feces are shown in grey, corn in blue and spinach in green.

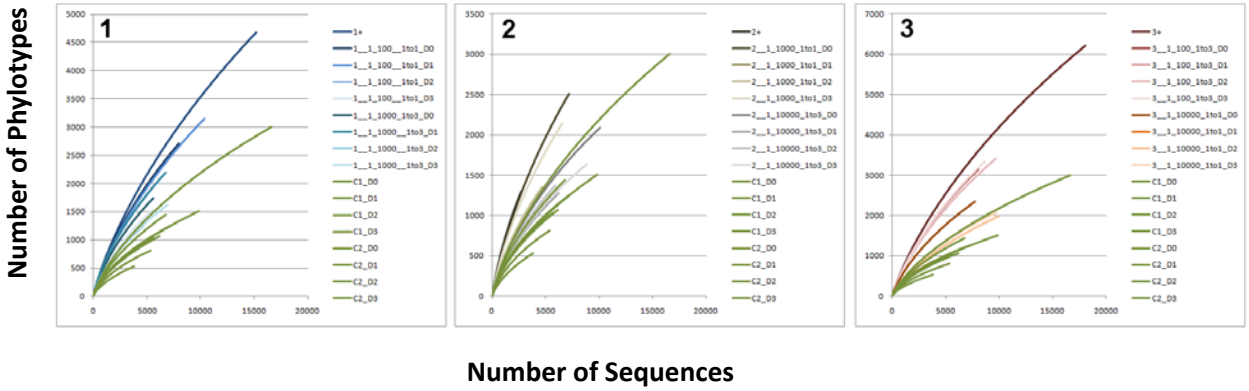

Figure 4 (1, 2, 3): Species richness of contaminated verse non-contaminated spinach samples. Diversity was analyzed using the RDP Pyro pipeline where samples underwent rarefaction analysis at from results obtained from complete cluster linkage at 0.97. Fecal samples 1 (blues), 2 (greys), and 3 (reds) were compared to two independent replicates of control samples (green) over a 3 day (D0, D1,D2,D3) period with each fecal sample dilution (1:100, 1:1000, 1:10000) and contamination ratio(1:1, 1:3) listed in the key (+) signifies a positive control where feces was added directly to the spinach without a dilution or contamination ratio. Figure 4-1 represents fecal sample 1, Figure 4-2 for fecal sample 2 and Figure 4-3 for fecal sample 3.

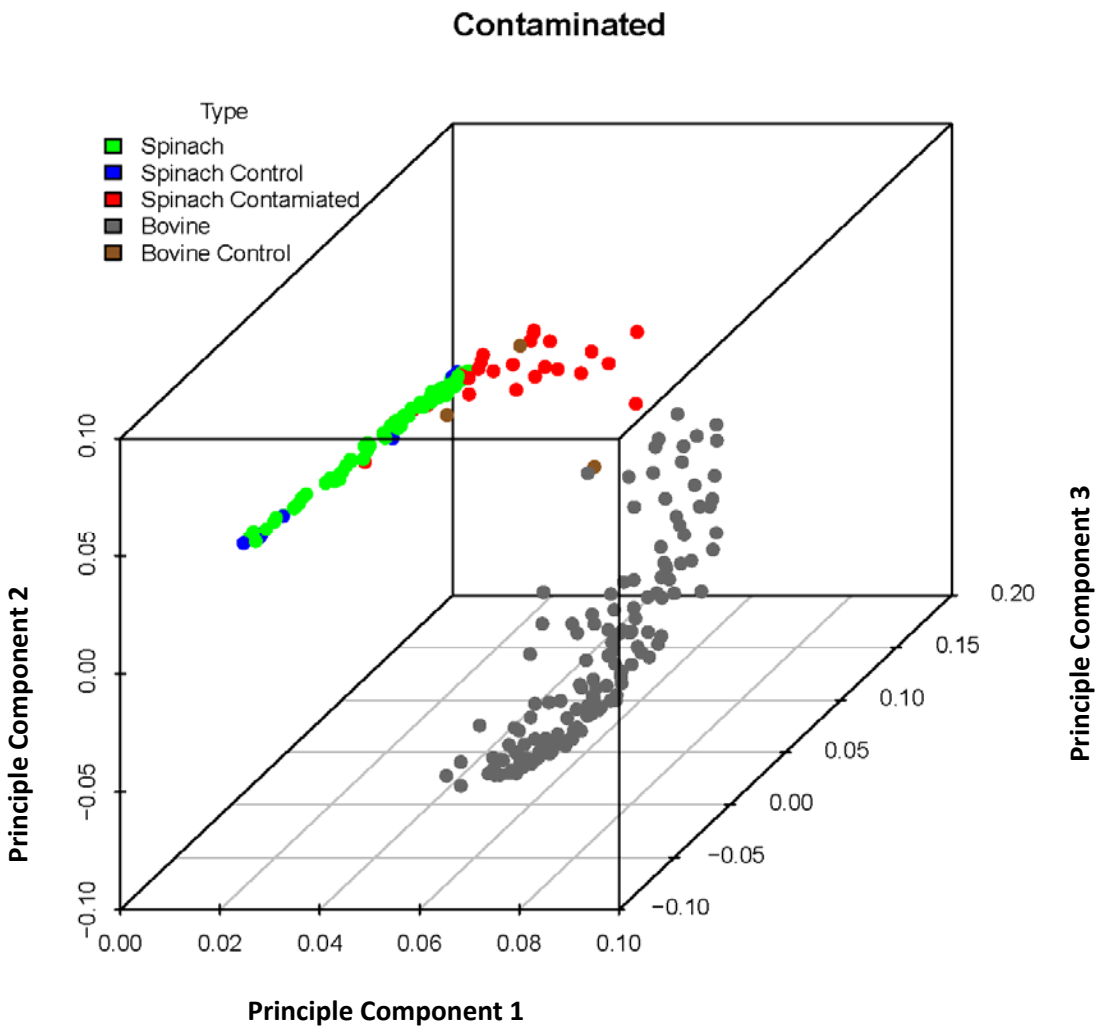

Figure 5 - PCA grouping of species level taxa between bovine feces, spinach leaves, and pooled fecal contaminated or uncontaminated spinach leaves . 424 species were used to conduct PCA analysis across all samples for the five environments above the minimum abundance threshold of 0.1% (Spinach Leaves, and Bovine Feces, Bovine Fecal Controls, Pooled Spinach Leaves Contaminated with Bovine Feces and Uncontaminated Pooled Spinach Leaves). The additive effects of species explaining the variation of these environments is primarily in the first three principal components (1 – 38.18%, 2 – 31.68%, 3 – 4.46%).

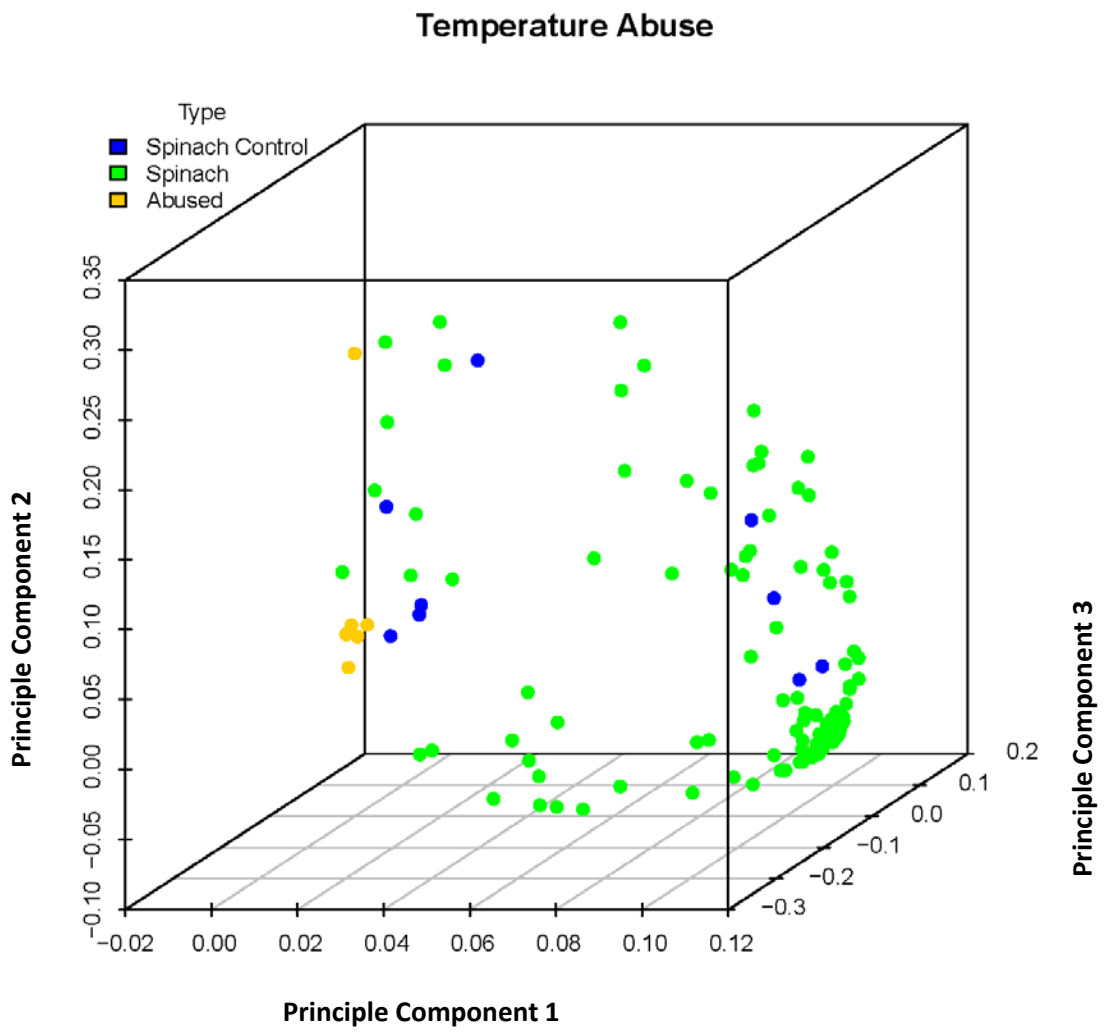

Figure 6 - PCA grouping of species level taxa between spinach leaves and pooled temperature abused or un-abused spinach leaves. 309 species were used to conduct PCA analysis across all samples for the three environments above the minimum abundance threshold of 0.1% (Spinach Leaves, Pooled Control Spinach Leaves and Pooled Abused Spinach Leaves). The additive effects of species explaining the variation of these environments is primarily in the first three principal components (1 – 64.49%, 2 – 7.52%, 3 – 5.06%).

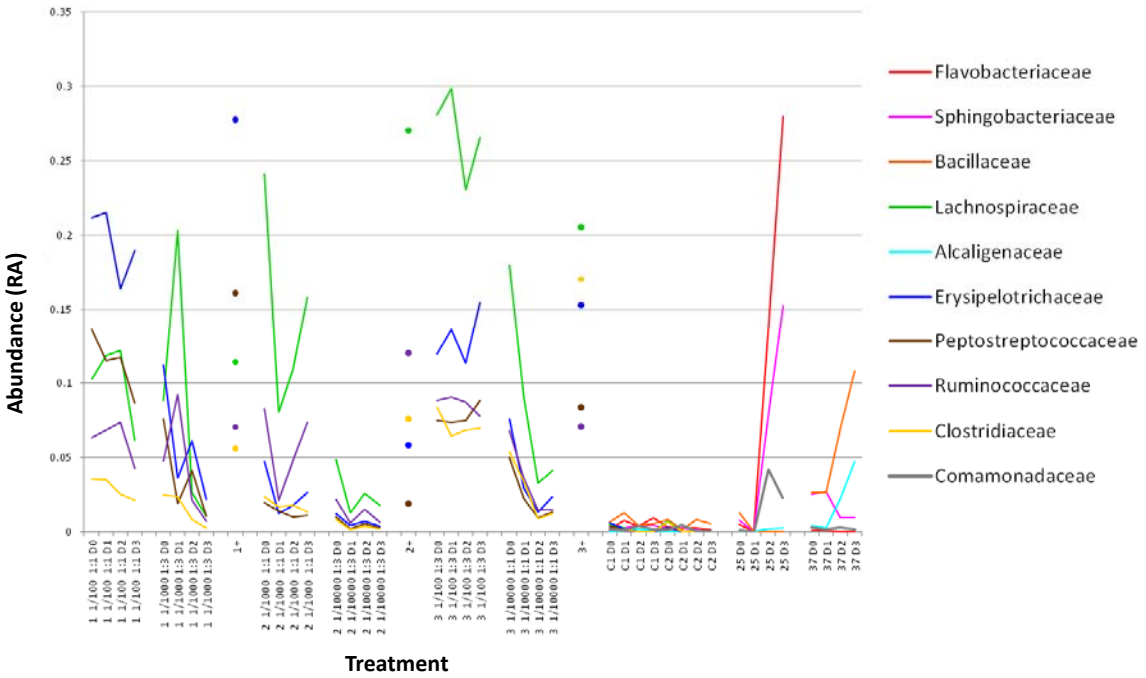

Figure 7 – Effects of contamination/abuse on the spinach epiphyte. (C) Significant Family level taxa were plotted by relative abundance along the 4 (D0,D1,D2,D3) day time-course. Taxa represented the three fecal samples and heat abused samples which are labeled according to abuse type (1,2,3) for contain fecal signatures at different dilutions (1:100, 1:1000, 1:10000) and ratios (1:1, 1:3) where (+) indicates positive fecal controls and (25,37) are temperature abused at the two different temperatures. C denotes negative spinach controls.

Use of pyrosequencing to identify new indicators of fecal contamination and temperature abuse in leafy greens.

Ryan Legge1,2, Jaehyoung Kim1,2, Chaomei Zhang3, Min Zhang, and Andrew K. Benson1,2,*

1Department of Food Science and Technology, and 2Core for Applied Genomics and Ecology, University of Nebraska, Lincoln, NE 68583-0919; 3 Department of Global Health, College of Public Health, University of South Florida, Tampa, FL 33612”

4. DISSCUSSION

After the seminal report in 2005 of the 454-based massively parallel sequencing (101). Pyrosequencing has been used to explore microbial communities as they relate to environmental microbiology, human health, and animal health (18-20, 50, 115, 184). In the work outlined through this dissertation we have further adopted pyrosequencing to two different projects that emphasize understanding how microbiomes assemble and influence physiological traits (heat loss and feed intake) and secondly on application of the technology to diagnostic microbiology. Both of these applications share the same fundamental aspect of microbiome assemblies and require emphasis on reliable approaches for classification and quantification of the data. In both studies, we emphasize use of parametric statistical methods. It is possible that non-parametric approaches would augment data analysis and allow the use of the full data sets. However, sample error will still be a problem for non-normally distributed data.

Despite the plurality of distributions that can be observed from multiple taxa, the dominant members of the microbiota from mice and the spinach phylloplane comprise a relatively small number of taxa and make up a large proportion of the biomass. Changes in these core sets of taxa were instrumental in identifying responses to selection in the mouse model and contamination in the spinach model. The ability to discern these changes was, of course, dependent upon significant numbers of samples to overcome the effect of sampling error and variation in the microbiota.

In addition to application of taxonomy-based parametric statistical analysis, we also used pattern-finding approaches such as PCA to differentiate samples and help

identify taxa contributing to the patterns. One aspect of this strategy is that the entire set of taxa was considered as opposed to treatment of taxa as discreet units in ANOVA. Use of PCA-type methods may therefore help identify natural trends in the data and be less subject to multiple testing errors associated with ANOVA.

Overall, these first explorations into the microbiome have taught us that as a phenotype or as an indicator, the complexity of the microbiome will demand application of robust statistical and analytical methodologies. Particularly for the indicator testing it will be unlikely that users could have the requisite expertise in the food industry. Therefore, robust automated methods for data analysis will continue to be a need.

1. Schidlowski M (2001) Carbon isotopes as biogeochemical recorders of life over 3.8 ga of earth history: Evolution of a concept. *Precambrian Research* 106: 117.
2. Carlson CJ, *et al* (2002) Effect of nutrient amendments on bacterioplankton production, community structure, and DOC utilization in the northwestern Sargasso Sea. *Aquatic Microbial Ecology* 30: 19.
3. Steinhoff U (2005) Who controls the crowd?. New findings and old questions about the intestinal microflora. *Immunol Lett* 99: 12.
4. O'Hara AM & Shanahan F (2006) The gut flora as a forgotten organ. *EMBO Rep* 7: 688.
5. Hutchinson GE (1961) The paradox of the plankton. *The American Naturalist* 95: 137.
6. Gause GF (1934) *The struggle for existence.* (Hafner Publishing Company, New York, NY),
7. Chase JM & Leibold MA (2003) *Ecological Niches: linking classical and contemporary approaches.* (The University of Chicago Press, London, UK),
8. Adler PB, HilleRisLambers J & Levine JM (2007) A niche for neutrality. *Ecol Lett* 10: 95.
9. Chesson P (2000) Mechanisms of maintenance of species diversity. *Annu Rev Ecol Syst* 31: 343.
10. Levine JM & HilleRisLambers J (2010) The maintenance of species diversity. *Nature Education Knowledge* 1: 67.
11. Yachi S & Loreau M (1999) Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *Proc Natl Acad Sci* 96: 1463.
12. Torsvik VL & Øvreås L (2011) in *Handbook of Molecular Microbial Ecology I,* (John Wiley & Sons, Inc.): 3.
13. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA & Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307: 1915.
14. McCann KS (2000) The diversity–stability debate. *Nature* 405: 228.
15. Zoetendal EG, Akkermans ADL, Akkermans-van Vliet WM, de Visser JAGM & de Vos WM (2001) The host genotype affects the bacterial community in the human gastrointestinal tract. *Micro Ecol Health Dis* 13: 129.
16. Benson AK, *et al* (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci* 107: 18933.
17. Qin J, *et al* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59.
18. Frank DN, *et al* (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci* 104: 13780.
19. Mazmanian SK, Round JL & Kasper DL (2008) A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* 453: 620.
20. McKenna P, *et al* (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* 4: e20.
21. Beres KA, Wallace RL & Segers HH (2005) Rotifers and Hubbel's unified neutral theory of biodiversity and biogeography. *Nat Resour Model* 18: 363.
22. Harpole W (2010) Neutral theory of species diversity. *Nature Education Knowledge* 1: 31.
23. Akkermans ADL, *et al* (1994) Molecular ecology of microbes: A review of promises, pitfalls and true progress. *FEMS Microbiology Reviews* 15: 185.
24. Stahl DA, Flesher B, Mansfield HR & Montgomery L (1988) Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. *Appl Environ Microbiol* 54: 1079.
25. Mack WN (1977) in *Bacterial Indicators: Health Hazards Associated with Water,* eds Hoadlye AW & Dutka BJ (American Society for Testing and Material, Tallahasse, FL), pp 59.

26. Amann RI, Ludwig W & Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59: 143.
27. Muyzer G & Smalla K (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek* 73: 127.
28. Torsvik V, Sørheim R & Goksøyr J (1996) Total bacterial diversity in soil and sediment communities—A review. *Journal of Industrial Microbiology & Biotechnology* 17: 170.
29. Hobbie JE, Daley RJ & Jasper J (1977) Use of nuclepore filters for counting bacteria by fluorescence microscopy. *Appl Environ Microbiol* 33: 1225.
30. Nannipieri P, Johnson RL & Paul EA (1978) Criteria for measurement of microbial growth and activity in soil. *Soil Biology and Biochemistry* 10: 223.
31. Karl DM (1979) Measurement of microbial activity and growth in the ocean by rates of stable ribonucleic acid synthesis. *Appl Environ Microbiol* 38: 850.
32. Eckburg PB, *et al* (2005) Diversity of the human intestinal microbial flora. *Science* 308: 1635.
33. Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 74: 5463.
34. Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology* 210: 1518.
35. Torsvik V, Goksøyr J & Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* 56: 782.
36. Rådström P, Knutsson R, Wolffs P, Lövenklev ,Maria & Löfström C (2004) Pre-PCR processing. *Mol Biotechnol* 26: 133.
37. Schneegurt MA, Dore SY & Kulpa CF (2003) Direct extraction of DNA from soils for studies in microbial ecology. *Curr Issues Mol Biol* 5: 1.
38. Yeates C, Gillings MR, Davison AD, Altavilla N & Veal DA (1998) Methods for microbial DNA extraction from soil for PCR amplification. *Biological Procedures Online* 1: 39.
39. Turnbaugh PJ, *et al* (2008) A core gut microbiome in obese and lean twins. . *Nature* 457: 480.
40. Ley RE, *et al* (2008) Evolution of mammals and their gut microbes. *Science* 320: 1647.
41. Martínez I, *et al* (2009) Diet-induced metabolic improvements in a hamster model of hypercholesterolemia are strongly linked to alterations of the gut microbiota. *Appl Environ Microbiol* 75: 4175.
42. Saiki R, *et al* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239: 487.
43. Gunn JS, Alpuche-Aranda CM, Loomis WP, Belden WJ & Miller SI (1995) Characterization of the *salmnella typhimurium pagC/pagD* Chromosomal region. . *J Bacteriol* 177: 5040.
44. Louis P & Flint HJ (2007) Development of a semiquantitative degenerate real-time PCR-based assay for estimation of numbers of butyryl-coenzyme A (CoA), CoA transferase genes in complex bacterial samples. *Appl Environ Microbiol* 73: 2009.
45. Widmer F, Shaffer BT, Porteous LA & Seidler RJ (1999) Analysis of nifH gene pool complexity in soil and litter at a Douglas fir forest site in the Oregon cascade mountain range. *Appl Environ Microbiol* 65: 374.
46. Demba Diallo M, Reinhold-Hurek B & Hurek T (2008) Evaluation of PCR primers for universal nifH gene targeting and for assessment of transcribed nifH pools in roots of oryza longistaminata with and without low nitrogen input. *FEMS Microbiol Ecol* 65: 220.
47. Rose TM, *et al* (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Research* 26: 1628.
48. Lane DJ, *et al* (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci* 82: 6955.
49. Ley RE, *et al* (2005) Obesity alters gut microbial ecology. . *Proc Natl Acad Sci* 102: 11070.
50. Sogin ML, *et al* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci* 103: 12115.

51. Clarridge JE (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews* 17: 840.
52. Murray PR, Baron EJ, Jorgensen JH, Landry ML & Pfaller MA (2007) *Manual of clinical microbiology,* (ASM Press, Washington, D.C. :),
53. Garrity GM, Bell JA & Lilburn TG (2004) *Taxonmic Outline of the Prokaryotes* (Springer-Verlang, New York), pp DOI:10.1007/bergeysoutline200405.
54. Woese C (1987) Bacterial evolution. *Microbiol Rev* 51: 221.
55. Woese C, Sogin M, Stahl D, Lewis BJ & Bonen L (1976) A comparison of the 16S ribosomal RNAs from mesophilic and thermophilic bacilli: Some modifications in the sanger method for RNA sequencing . *Journal of Molecular Evolution* 7:197.
56. Olsen GJ, Lane DJ, Giovannoni SJ & Pace NR (1986) Microbial ecology and evolution: A ribosomal RNA approach. *Annu Rev Microbiol* 40: 337.
57. Lauber CL, Hamady M, Knight R & Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75: 5111.
58. Samaha RR, O'Brien B, O'Brien TW & Noller HF (1994) Independent in vitro assembly of a ribonucleoprotein particle containing the 3' domain of 16S rRNA. . *Proc Natl Acad Sci* 91: 7884.
59. Li G, Oh E & Weissman JS (2012) The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484: 538.
60. Lane DJ (1991) in *Nucleic acid techniques in bacterial systematics*. eds Stackebrandt E & Goodfellow M (John Wiley & Sons, Chichester, United Kingdom), pp 115.
61. Wang Q, Garrity GM, Tiedje JM & Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73: 5261.
62. Baker GC, Smith JJ & Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* 55: 541.
63. Morales SE & Holben WE (2009) Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specify and efficacy not suggested by in silico analysis. *Appl Environ Microbiol* 75:2677.
64. Case RJ, *et al* (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* 73: 278.
65. Farrelly V, Rainey FA & Stackebrandt E (1995) Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol* 61: 2798.
66. Ward DM, Bateson MM, Weller R & Ruff-Roberts AL (1992) Ribosomal RNA analysis of microorganisms as they occur in nature. *Adv Microb Ecol* 12: 219.
67. Jin T, Zhang T & Yan Q (2010) Characterization and quantification of ammonia-oxidizing archaea (AOA) and bacteria (AOB) in a nitrogen-removing reactor using T-RFLP and qPCR. *Appl Microbiol Biotechnol* 87: 1167.
68. Osborn AM, Moore ERB & Timmis KN (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ Microbiol* 2: 39.
69. Lukow T, Dunfield PF & Liesack W (2000) Use of the T-RFLP technique to assess spatial and temporal changes in the bacterial community structure within an agricultural soil planted with transgenic and non-transgenic potato plants. *FEMS Microbiol Ecol* 32: 241.
70. Nagashima K, Hisada T, Sato M & Mochizuki J (2003) Application of new primer-enzyme combinations to terminal restriction fragment length polymorphism profiling of bacterial populations in human feces. *Appl Environ Microbiol* 69: 1251.
71. Hamady M, Walker JJ, Harris JK, Gold NJ & Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Meth* 5: 235.

72. Lozupone CA & Knight R (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci* 104: 11436.
73. Walter J, *et al* (2000) Detection and identification of gastrointestinal *Lactobacillus* species by using denaturing gradient gel electrophoresis. *Appl Environ Microbiol* 66: 297.
74. Nocker A, Burr M & Camper A (2007) Genotypic microbial community profiling: A critical technical review. *Microb Ecol* 54: 276.
75. Avaniss-Aghajani E, *et al* (1996) Molecular technique for rapid identification of mycobacteria. *J Clin* 34: 98.
76. Liu WT, Marsh TL, Cheng H & Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of gene encoding 16S rRNA. *Appl Environ Microbiol* 63: 4516.
77. Egert M & Friedrich MW (2003) Formation of pseudo-terminal restriction fragments, a PCR-related bias affecting terminal restriction fragment length polymorphism analysis of microbial community structure. *Appl Environ Microbiol* 69: 2943.
78. Reysenbach AL, Giver LJ, Wickham GS & Pace NR (1992) Differential amplification of rRNA genes by polymerase chain reaction. *Appl Environ Microbiol* 58: 3417.
79. Fierer N, Jackson JA, Vilgalys R & Jackson BB (2005) Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. *Appl Environ Microbiol* 71: 4117.
80. Campbell BJ, Yu L, Heidelberg JF & Kirchman DL (2011) Activity of abundant and rare bacteria in a coastal ocean. . *Proc Natl Acad Sci* 108: 12776.
81. Sharma S, *et al* (2007) Quantification of functional genes from prokaryotes in soil by PCR. *J Microbiol Methods* 68: 445.
82. Baldwin BR, Nakatsu CH & Nies L (2003) Detection and enumeration of aromatic oxygenase genes by multiplex and real-time PCR. *Appl Environ Microbiol* 69: 3350.
83. Heid CA, Stevens J, Livak KJ & Williams PM (1996) Real time quantitative PCR. *Genome Res* 6: 986.
84. Radajewski S, Ineson P, Parekh NR & Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* 403: 646.
85. Meselson M & Stahl FW (1958) The replication of DNA in *Escherichia coli*. . *Proc Natl Acad Sci* 44: 671.
86. Radajewski S, Ineson P, Parekh NR & Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* 403: 646.
87. Sessitsch A, *et al* (2006) Diagnostic microbial microarrays in soil ecology. *New Phytol* 171: 719.
88. Sagaram US, *et al* (2009) Bacterial diversity of huanglogbing pathogen-infected citrus using PhyloChips and 16S rDNA clone library sequencing. *Appl Environ Microbiol* 75: 1566.
89. Xu J (2011) in *Handbook of Molecular Microbial Ecology I*, (John Wiley & Sons, Inc.): 111.
90. Bottari B, Ercolini D, Gatti M & Neviani E (2006) Application of FISH technology for microbiological analysis: Current state and prospects. *Appl Micoribal Biotech* 73: 485.
91. Levsky J & Singer R (2003) Fluorescence in situ hybridization: Past, present and future. *J Cell Sci* 116: 2833.
92. Amann RI, Ludwig W & Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* 59: 143.
93. Meier H, Amann R, Ludwig W & Schleifer KH (1999) Specific oligonucleotide probes for in situ detection of a major group of gram-positive bacteria with low DNA G + C content. *Syst Appl Microbiol* 22: 186.
94. Roller C, Wagner M, Amann R, Ludwig W & Schleifer KH (1994) In situ probing of gram-positive bacteria with high DNA G + C content using 23S rRNA-targeted oligonucleotides. *Microbiology* 140: 2849.

95. Fuchs BM, *et al* (1998) Flow cytometric analysis of the in situ accessibility of *Escherichia coli* 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl Environ Microbiol* 64: 4973.
96. Blattner FR, *et al* (1997) The complete genome sequence of *Escherichia coli K-12*. *Science* 277: 1453.
97. Fleischmann RD, *et al* (1995) Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science* 269: 496.
98. Lander ES, *et al* (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860.
99. Weinstock GM (2011) in *Handbook of Molecular Microbial Ecology I*, (John Wiley & Sons, Inc.): 143.
100. Giovannoni SJ, Britschgi TB, Moyer CL & Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345: 60.
101. Margulies M, *et al* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.
102. Harismendy O, *et al* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 10: R32.
103. Zhou X, *et al* (2010) The next-generation sequencing technology and application. *Protein & Cell* 1: 520.
104. Metzker ML (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics* 11: 31.
105. J. W & Ansorge (2009) Next-generation DNA sequencing techniques. *New Biotechnology* 25: 195.
106. Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Meth* 5: 16.
107. Shendure J & Ji H (2008) Next-generation DNA sequencing. *Nat Biotech* 26: 1135.
108. Sula WJ, *et al* (2011) Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proc Natl Acad Sci* 108: 14637.
109. Dowd SE, *et al* (2008) Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiology* 8: 125.
110. Turnbaugh PJ, *et al* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027.
111. Hamady M & Kinght R (2009) Microbial community profiling for human microbiome project: Tools, techniques, and challenges. *Genome Res* 19: 1141.
112. Luo C, Tsementzi D, Kyrpides NC & Konstantinidis KT (2011) Individual genome assembly from complex community short-read metagenomic datasets. *ISME J* 6:898.
113. Begley M & Hill C (2010) Food safety: What can we learn from genomics?. *Annu Rev Food Sci Technol* 1: 341.
114. Critzer FJ & Doyle MP (2010) Microbial ecology of foodborne pathogens associated with produce. *Curr Opin Biotechnol* 21: 125.
115. Lauber CL, Hamady M, Knight R & Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75: 5111.
116. Huse S, Huber J, Morrison H, Sogin M & Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 8: R143.
117. Kunin V, Engelbrektson A, Ochman H & Hugenholtz P (2010) Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 12: 118.
118. Quince C, *et al* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Meth* 6: 639.

119. Gomez-Alvarez V, Teal TK & Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3: 1314.
120. Engelbrektson A, *et al* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* 4: 642.
121. Zhou J, *et al* (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* 5: 1303.
122. Gill SR, *et al* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312: 1355.
123. Venter JC, *et al* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66.
124. Tyson GW, *et al* (2) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: 37.
125. García-Martín H, *et al* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* 10: 1263.
126. Bodrossy L & Sessitsch A (2004) Oligonucleotide microarrays in microbial diagnostics. *Current Opinion in Microbiology* 7: 245.
127. Simon C & Daniel R (2009) Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol* 85: 265.
128. Arumugam M, *et al* (2011) Enterotypes of the human gut microbiome. *Nature* 473: 174.
129. McDonald D, *et al* (2011) An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6:610.
130. Pruesse E, *et al* (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 35: 7188.
131. Cole JR, *et al* (2009) The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucl Acids Res* 37 (suppl 1): D141.
132. Cole JR, Wang Q, Chai B & Tiedje JM (2011) in *Handbook of Molecular Microbial Ecology I,* (John Wiley & Sons, Inc.): 313.
133. Ludwig W, *et al* (2004) ARB: A software environment for sequence data. *Nucleic Acids Research* 32: 1363.
134. Thomas T, Gilbert J & Meyer F (2012) Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2: 3.
135. Glass EM, Wilkening J, Wilke A, Antonopoulos D & Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols* 2010: pdb.prot5368.
136. Overbeek R, *et al* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* 33: 5691.
137. Glass EM & Meyer F (2011) in *Handbook of Molecular Microbial Ecology I,* (John Wiley & Sons, Inc.): 325.
138. Huson DH, Auch AF, Qi J & Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research* 17: 377-386.
139. Tatusov R, *et al* (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.
140. Huson DH & Mitra S (2011) in *Handbook of Molecular Microbial Ecology I,* (John Wiley & Sons, Inc.): 352.
141. Alterovitz G, Xiang M, Mohan M & Ramoni MF (2007) GO PaD: The gene ontology partition database. *Nucleic Acids Research* 35: D322.
142. Johnson M, *et al* (2008) NCBI BLAST: A better web interface. *Nucl Acids Res* 36: W5.
143. Li W & Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658.

144. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460.
145. Salzberg SL, Sommer DD, Puiu D & Lee VT (2008) Gene-boosted assembly of a novel bacterial genome from very short reads. *PLoS Comput Biol* 4: e1000186.
146. Dutilh BE, Huynen MA, Gloerich J & Strous M (2011) in *Handbook of Molecular Microbial Ecology I*, (John Wiley & Sons, Inc.): 385.
147. Li R, Li Y, Kristiansen K & Wang J (2008) SOAP: Short oligonucleotide alignment program. *Bioinformatics* 24: 713.
148. Zerbino DR & Birney E (2008) Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research* 18: 821.
149. Peng Y, Leung HCM, Yiu SM & Chin FYL (2011) Meta-IDBA: A de novo assembler for metagenomic data. *Bioinformatics* 27: i94.
150. Martin-Laurent F, *et al* (2001) DNA extraction from soils: Old bias for new microbial diversity analysis methods. *Appl Environ Microbiol* 67: 2354.
151. Luna GM, Dell'Anno A & Danovaro R (2006) DNA extraction procedure: A critical issue for bacterial diversity assessment in marine sediments. *Environmental Microbiology* 8: 308.
152. Wang GC & Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. . *Appl Environ Microbiol* 63: 4645.
153. Polz MF & Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* 68: 3724.
154. Stackebrandt E & Goebel BM (1994) Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology . *Int J Syst Bacteriol* 44: 846.
155. Bergthorsson U & Ochman H (1998) Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution* 15: 6.
156. Philippe H & Douady CJ (2003) Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol* 6: 498.
157. Wayne LG, *et al* (1987) International committee on systematic bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic Bacteriology* 37: 463.
158. Konstantinidis KT & Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci* 102: 2567.
159. Curtis T, Sloan WT & Scannell J (2002) Modelling prokaryotic diversity and its limits. *Proc Natl Acad Sci* 99: 10494.
160. Roesch LFW, *et al* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* 1: 283.
161. Rosenberg E, Sharon G, Atad I & Zilber-Rosenberg I (2010) The evolution of animals and plants via symbiosis with microorganisms. *Environmental Microbiology Reports* 2: 500-506.
162. Shively JM, English RS, Baker SH & Cannon GC (2001) Carbon cycling: The prokaryotic contribution. *Curr Opin Microbiol* 4: 301.
163. Maloy S & Schaechter M (2006) The era of microbiology: A golden phoenix. *Int Microbiol* 9:1.
164. Singh S, Singh B, Mishra BK, Pandey AK & Nain L (2012) *in Microorganisms in Sustainable Agriculture and Biotechnology* (Springer-Netherlands), pp. 127.
165. Bardgett RD, Freeman C & Ostle NJ (2008) Microbial contributions to climate change through carbon cycle feedbacks. *ISME J* 2: 805.
166. Jannasch HW & Mottl MJ (1985) Geomicrobiology of deep-sea hydrothermal vents. *Science* 229: 717.
167. Corliss JB, *et al* (1979) Submarine thermal sprirngs on the Galápagos rift. *Science* 203: 1073.

168. Oremland RS & Stolz JF (2003) The ecology of arsenic. *Science* 300: 939.
169. Leroy F & Vuyst LD (2004) Lactic acid bacteria as functional starter cultures for the food fermentation industry. *Trends Food Sci Technol* 15: 67.
170. H. K & Steinkraus (1994) Nutritional significance of fermented foods. *Food Res Int* 27: 259.
171. Ghisalba O (1983) Microbial degradation of chemical waste, an alternative to physical methods of waste disposal. *Cellular and Molecular Life Sciences* 39: 1247.
172. Kumar R, Verma D, Singh BL, Kumar U & Shweta (2010) Composting of sugar-cane waste by-products through treatment with microorganisms and subsequent vermicomposting. *Bioresour Technol* 101: 6707.
173. Sayavedra-Soto L, Gvakharia B, Bottomley P, Arp D & Dolan M (2010) Nitrification and degradation of halogenated hydrocarbons - tenuous balance for ammonia-oxidizing bacteria. *Appl Microbiol Biotechnol* 86: 435.
174. Figueroa-Gonzelez I, Quijano G, Ramerez G & Cruz-Guerrero A (2011) Probiotics and prebiotics-perspectives and challenges. *J Sci Food Agric* 91: 1341.
175. Gaggia F, Mattarelli P & Biavati B (2010) Probiotics and prebiotics in animal feeding for safe food production. *Int J Food Microbiol* 141, Supplement: S15.
176. Patterson J & Burkholder K (2003 Apr) Application of prebiotics and probiotics in poultry production. *Poultry Science* 82: 627.
177. Roberfroid MB (2000) Prebiotics and probiotics: Are they functional foods?. *The American Journal of Clinical Nutrition* 71: 1682S.
178. Hughes-Martiny JB, *et al* (2006) Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiol* 4: 102.
179. Savage DC (1977) Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* 31: 107.
180. Walter J, Britton RA & Roos S (2011) Host-microbial symbiosis in the vertebrate gastrointestinal tract and the *Lactobacillus reuteri* paradigm. *Proc Natl Acad Sci* 108: 4645.
181. Oh PL, *et al* (2009) Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *ISME J* 4: 377.
182. Frese SA, *et al* (2011) The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet* 7: e1001314.
183. Caporaso JG, *et al* (2011) Moving pictures of the human microbiome. *Genome Biology* 12: R50.
184. Ley RE, Turnbaugh PJ, Klein S & Gordon JI (2006) Microbial ecology: Human gut microbes associated with obesity. *Nature* 444: 1022.
185. Stevens CE & Hume ID (1998) Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients. *Physiol Rev* 78: 393.
186. Sela DA, *et al* (2008) The genome sequence of *Bifidobacterium longum subsp. infantis* reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci* 105: 18964.
187. Ochman H, *et al* (2010) Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol* 8: e1000546.
188. Ley RE, Lozupone CA, Hamady M, Knight R & Gordon JI (2008) Worlds within worlds: Evolution of the vertebrate gut microbiota . *Nat Rev Microbiol* 6: 776.
189. Vijay-Kumar M, *et al* (2010) Metabolic syndrome and altered gut microbiota in mice lacking toll-like receptor 5. *Science* 328: 228.
190. Rehman A, *et al* (2011) Nod2 is essential for temporal development of intestinal microbial communities. *Gut* 60: 1354.
191. Nielsen MK, *et al* (1997b) Divergent selection for heat loss in mice: II. Correlated responses in feed intake, body mass, body composition, and number born through fifteen generations. *J Anim Sci* 75: 1469.

192. Nielsen MK, Jones LD, Freking BA & DeShazer JA (1997a) Divergent selection for heat loss in mice: I. Selection applied and direct response through fifteen generations *J Anim Sci* 75: 1461.
193. McDonald JM & Nielsen MK (2007) Renewed selection for heat loss in mice: Direct responses and correlated responses in feed intake, body weight, litter size, and conception rate. *J Anim Sci* 85: 658.
194. Eggert DL & Nielsen MK (2006) Costs of lean deposition, fat deposition and maintenance in three lines of mice selected for heat loss. *J Anim Sci* 84: 276.
195. Mousel MR, Stroup WW & Nielsen MK (2001) Locomotor activity, core body temperature, and circadian rhythms in mice selected for high or low heat loss. *J Anim Sci* 79: 861.
196. Peterson DA, McNulty NP, Guruge JL & Gordon JI (2007) IgA response to symbiotic bacteria as a mediator of gut homeostasis. *Cell Host & Microbe* 2: 328.
197. Hooper LV, *et al Methods in Microbiology,* (Academic Press): 559.
198. Pruesse E, *et al* (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res* 35: 7188.
199. Zar JH (2005) in *Encyclopedia of Biostatistics,* (John Wiley & Sons, Inc.)
200. Kluger MJ, Conn CA, Franklin B, Freter R & Abrams GD (1990) Effects of gastrointestinal flora on body temperature of rats and mice. *Am J Physiol Regulatory Integrative Comp Physiol* 258: 553.
201. Dohi T, *et al* (2004) $CD4^{+}CD45RB^{HI}$ interleukin-4 defective T cells elicit antral gastritis and duodenitis. *AJP* 165: 1257.
202. Tlaskalova-Hogenova H, *et al* (2011) The role of gut microbiota (commensal bacteria) and the mucosal barrier in the pathogenesis of inflammatory and autoimmune diseases and cancer: Contribution of germ-free and gnotobiotic animal models of human diseases. *Cell Mol Immunol* 8: 110.
203. Wostmann BS, Larkin C, Moriarty A & Bruckner-Kardoss E (1983) Dietary intake, energy metabolism, and excretory losses of adult male germfree Wistar rats. *Lab Anim Sci* 33: 46.
204. Bäckhed F, *et al* (2004) The gut microbiota as an environmental factor that regulates fat storage. . *Proc Natl Acad Sci* 101: 15718.
205. Sanz Y, Santacruz A & De Palma G (2008) Insights into the roles of gut microbes in obesity. *Interdisciplinary Perspectives on Infectious Diseases* 2008:
206. Swallow JG & Garland T,Jr (2005) Selection experiments as a tool in evolutionary and comparative physiology: Insights into complex traits--an introduction to the symposium . *Integr Comp Biol* 45: 387.
207. Garland T,Jr, *et al* (2011) How to run far: Multiple solutions and sex-specific responses to selective breeding for high voluntary activity levels . *Proc Biol Sci* 278: 574.
208. Osawa R, Fujisawa T & Pukall R (2006) *Lactobacillus apodemi sp. nov.,* a tannase-producing species isolated from wild mouse feces. *International Journal of Systematic and Evolutionary Microbiology* 56: 1693.
209. Sasaki E, *et al* (2005) Isolation of tannin-degrading bacteria isolated from feces of the Japanese large wood mouse, *Apodemus speciosus*, feeding on tannin-rich acorns. *Syst Appl Microbiol.*28: 358.
210. Rhoades DF & Cates RG (1976) Toward a general theory of plant antiherbivore chemistry. *Rec Adv Phytochem* 10: 168.
211. Lindroth RL & Batzli GO (1984) Plant phenolics as chemical defenses: Effects of natural phenolics on survival and growth of prairie voles (*Microtus ochrogaster*). *J Chem Ecol* 10: 229.
212. Cooper SM & Owen-Smith N (1985) Condensed tannins deter feeding by browsing ruminants in a South African savanna. *Oecologia* 67: 142.
213. Robbins CT, *et al* (1987) Role of tannins in defending plants against ruminants: Reduction in protein availability. *Ecology* 68: 98.

214. Robbins CT, Mole S, Hagerman AE & Hanley TA (1987) Role of tannins in defending plants against ruminants: Reduction in dry matter digestion?. *Ecology* 68: pp. 1606-1615.
215. Centers for Disease Control and Prevention (2011) CDC estimates of foodborne illness in the united states2011: 2.
216. Gould LH, *et al* (2011) Surveillance for foodborne disease outbreaks --- United States, 2008. *MMWR* 60: 1197.
217. Ayers LT, Williams IT, Grey S, Griffin PM & Hall AJ (2009) Surveillance for foodborne disease outbreaks --- United States, 2006. *MMWR* 58: 609.
218. Boore A, *et al* (2010) Surveillance for foodborne disease outbreaks --- United States, 2007. *MMWR* 59: 973.
219. Feng P, Weagant SD & Grant MA (2002) in *Bacteriological Analytical Manual (BAM),* eds Hammack T, et al (U.S. Food and Drug Administration).
220. Schloss PD & Handelsman J (2005) Metagenomics for studying unculturable microorganisms: Cutting the Gordian knot. *Genome Biol* 6: 229.
221. Amann RI, *et al* (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol* 56: 1919.
222. Colwell RR, *et al* (1985) Viable but non-culturable *Vibrio cholerae* and related pathogens in the environment: Implications for release of genetically engineered microorganisms. *Nature Biotechnology* 3: 817.
223. Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* 88: 1354.
224. Leclerc H, Mossel DA, Edberg SC & Struijk CB (2001) Advances in the bacteriology of the coliform group: Their suitability as markers of microbial water safety. *Annu Rev Microbiol* 55: 201.
225. Berg G (1978) in *Indicators of Viruses in Water and Food,* (Ann Arbor Science, Ann Arbor, MI): 424.
226. Van Poucke SO & Nelis HJ (1997) Limitations of highly sensitive enzymatic presence-absence tests for detection of waterborne coliforms and *Escherichia coli. Appl Environ Microbiol* 63: 771.
227. Miskimin DK, *et al* (1976) Relationships between indicator organisms and specific pathogens in potentially hazardous foods. *J Food Sci* 41: 1001.
228. Nugen SR & Baeumne AJ (2008) Trends and opportunities in food pathogen detection. *Anal Bioanal Chem* 391: 451.
229. Kothary MH & Babu US (2001) Infective dose of foodborne pathogen in volunteers. *Journal of Food Saftey* 21: 49.
230. Rangel JM, Sparling PH, Crowe C, Griffin PM & Swerdlow DL (2005) Epidemiology of *Escherichia coli O157:H7* outbreaks, united states, 1982-2002. *Emerg Infec Dis* 11: 603.
231. Yu J, Holland JB, McMullen MD & Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178: 539.
232. McMullen MD, *et al* (2009) Genetic properties of the maize nested association mapping population. *Science* 325: 737.
233. Gore MA, *et al* (2009) A first-generation haplotype map of maize. *Science* 326: 1115.
234. Chassy B (2010) Can –omics inform a food safety assessment? . *Regul Toxicol Pharmacol* 58: S62.
235. Rohde H, *et al* (2011) Open-source genomic analysis of shiga-toxin producing *E. coli O104:H4. N Engl J Med* 365: 718.
236. Schellenberger S, Drake HL & Kolb S (2011) Functionally redundant cellobiose-degrading soil bacteria respond differentially to oxygen. *Appl Environ Microbiol* 77: 6043.
237. Nava GM, Friedrichsen HJ & Stappenbeck TS (2011) Spatial organization of intestinal microbiota in the mouse ascending colon. *The ISME Journal* 5: 627.

238. Kondoa R, Nedwella DB, Purdyb KJ & Silvaa SQ (2004) Detection and enumeration of sulphate-reducing bacteria in estuarine sediments by competitive PCR. *Geomicrobiology* 21: 145.
239. Dowd S, *et al* (2008) Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiology* 8: 125.
240. de Menezes AB, *et al* (2011) Microbiome analysis of dairy cows fed pasture or total mixed ration diets. *FEMS Microbiol Ecol* 78: 256.
241. Rudi K, Moen B, Sekelja M, Frisli T & Lee MRF (2012) An eight-year investigation of bovine livestock fecal microbiota. *Vet Microbiol* Online.
242. Bernardet J & Nakagawa Y (2006) An introduction to the family flavobacteriaceae455.
243. Babic I & Watada AE (1996) Microbial populations of fresh-cut spinach leaves affected by controlled atmospheres. *Postharvest Biol Technol* 9: 187.
244. Sun YZ, Mao SY & Zhu WY (2010) Rumen chemical and bacterial changes during stepwise adaptation to a high-concentrate diet in goats. *Animal* 4: 210.
245. Gong J, *et al* (2007) 16S rRNA gene-based analysis of mucosa-associated bacterial community and phylogeny in the chicken gastrointestinal tracts: From crops to ceca. *FEMS Microbiol Ecol* 59: 147.
246. Shanks OC, *et al* (2011) Community structures of fecal bacteria in cattle from different animal feeding operations. *Appl Environ Microbiol* 77: 2992.
247. Ternström A, Lindberg A- & Molin G (1993) Classification of the spoilage flora of raw and pasteurized bovine milk, with special reference to pseudomonas and bacillus. *J Appl Microbiol* 75: 25.
248. Tournas VH (2005) Spoilage of vegetable crops by bacteria and fungi and related health hazards. *Crit Rev Microbiol* 31: 33.