USING BIOINFORMATICS TO ANALYZE THE ROLE OF MICROBIAL TAXA IN COMPLEX ECOSYSTEMS

by

Nina Sanapareddy

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology
Charlotte
2011

Approved by:

_____
Dr. Anthony A. Fodor

_____
Dr. Christine Richardson

_____
Dr. Shannon Schlueter

_____
Dr. Zhengchang Su

ABSTRACT

NINA SANAPAREDDY. Using bioinformatics to analyze the role of microbial taxa in complex ecosystems. (Under the direction of DR. ANTHONY A. FODOR)

Microbes are abundant on earth and play a crucial role in the environment they inhabit. Before the dawn of metagenomics, the study of the effect of microorganisms on their environment was limited due to use of low throughput techniques that could only examine single organisms or a few at a time. Metagenomics is a fast growing field of science that permits investigation of microbes by directly extracting DNA from the environment. A lot of environments, ranging in complexity from the ocean to acid mines, from wastewater communities to the human body have been targeted by metagenomics studies, and these studies generate tremendous amounts of data and newer and more efficient bioinformatic tools and methods are needed to interpret this complex data.

In this dissertation we used bioinformatic tools to enrich our understanding of the role that microorganisms play within some important but understudied microbial environments. In Chapter # 1, we report an increased microbial richness associated with colorectal cancer. This is an important finding that could lead to the development of diagnostic methods to identify individuals at high risk of developing colorectal cancer and this early detection could help devise preventive strategies. In Chapter # 2 we discuss a batch-effect we discovered in our colorectal cancer project and how filtering out the batch-effect helped us in revealing the true biological signal. In Chapter #3 we report results of a metagenomic survey where we analyzed the pyrosequences obtained from a wastewater community. In Chapter # 4 of this dissertation we perform a systematic

comparison of some of the methods used in taxonomic profiling of microbial communities and show how the choice of method can have an effect on a community's taxonomic profile.

Overall, this dissertation demonstrates the value of using bioinformatic tools during the course of analysis of complex communities, in not only filtering out artifacts and in choice of analysis pathways but also in discovering important biological effects.

DEDICATION

This thesis is dedicated to my parents, who taught me the persistence and tenacity that I needed to fulfill my ambition. Without their unconditional support and encouragement, my goal of completing doctoral level education would've been difficult to achieve.

ACKNOWLEDGEMENTS

To start with, I would like to acknowledge the funding support from NIH; through grants: P30DK034987, R01 CA44684, P50 CA106991, R01 CA136887, and K01 DK 073695; for the colorectal cancer project, which gave rise to the Chapters 1 and 2 of this dissertation.

In addition, I am grateful to Dr Temitope O Keku, Associate Professor, UNC Chapel Hill, for being generous enough to allow us to analyze the very interesting colorectal cancer dataset. I appreciate Jon Mccafferty for helping me find chimeric sequences in the dataset described in Chapter 1 and Timm Hamp for his reliable and remarkable molecular biology skills in generating the wastewater dataset that was discussed in Chapters 3 and 4 of this dissertation. I would also like to thank everyone in my research group for providing a very productive, stimulating and pleasant environment to work in.

I am greatly indebted to my advisor Dr Anthony A. Fodor, for imparting his invaluable bioinformatics expertise to me and for providing the training, guidance, advice and direction throughout the course of my research. Finally, I would like to express my gratitude to my friends and family who have been a constant source of encouragement and support during this very important and demanding period of my life.

INTRODUCTION

The significance of microbes in the environment they inhabit: Microbes are everywhere and their presence always affects the environment that they are growing in. Microbes are found in almost every habitat on earth, ranging from extreme climates like acidic hot springs [1], radioactive waste [2], and Earth's crust [3] to relatively moderate ones like inside and on the surface of the plants and animal bodies [4]. Nearly all animals, plants and certain types of fungi are dependent on microbes because the microbes make vital minerals, nutrients and vitamins accessible to their hosts[5]. Microbes inhabit animal digestive systems, their mouths, their skin and many other organs and are important for the maintaining the health of their animal hosts. Comparisons of germ free mice with those colonized with microbiota[6], have shown that the microbiota help regulate energy balance, not only by extracting calories from otherwise indigestible components of our diet but also by controlling host genes that help in storage of the extracted energy. These studies thus conclude that manipulating the microbial composition may be helpful in regulating the energy balance in the hosts [7],[8],[9].

The role that microorganisms play in their environment has been a central focus of microbiology for a long time. However, in the past, microbiology focused on isolating one or a few species at a time, by culturing them individually, so very little insight was gained about all the members of that community, as a whole. Metagenomics, sequencing of DNA extracted directly from environmental samples is a new tool that helps us study microbes, not as separate entities but as a whole, in complex communities. Metagenomic studies on a wide variety of environments including the ocean, soil, thermal vents, acid

mine drainages and the human microbiome are helping to reveal the vast microbial diversity that has been hidden from us in the past due the limitations of the preexisting technologies[10], [11], [12]. Metagenomics has rapidly advanced in the recent past and this growth can be attributed, not only to the technical and analytical methods developed from high throughput platforms but also to the simultaneous advancements in the associated bioinformatics and statistical software [13-14].

Metagenomic analysis of microbial communities: The term “metagenome”[15] was coined by Jo Handelsman and was initially used to describe a collection of genes, from a number of genomes, sequenced directly from the environment that could be analyzed in the same way as a single genome. Recently though, metagenomics is being used in a broader sense, to describe any sequencing of genetic material from uncultured environmental samples, whether it is from an entire community, a single organism, all the genes or just one gene (like the 16S rRNA gene). Kevin Chen and Lior Pachter (researchers at the University of California, Berkeley) defined metagenomics as "the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species."[16].

The quality and quantity of results obtained from a metagenomic analysis of any community will be dependent upon the procedures used for sampling the community, on the molecular biological methods like DNA extraction on the sequencing methods used, and on the bioinformatic and statistical analytical methods used. Deciding on the best way to sample a microbial community for metagenomics is one of the biggest challenges faced in the planning phase of any metagenomic study. Time-course studies gauge the

response of the inhabitant microbes to changing conditions over time. These studies lead to a better understanding of the overall community structure, function, and its robustness to the changing conditions. Similarly, to comprehend the role of host-associated microbial communities in host development and health requires not only sampling from the same host over time (longitudinal studies), but also assessing host-to-host variation at a given point of time (cross-sectional and case-control studies). Habitat and host variability add more levels of complexity the already complex sampling related issues. Another source of variability, which is crucial in metagenomics studies, is technical variability. In studies involving large sample sizes, sometimes the samples are processed in batches and the quality of the data will depend on ensuring that, as far as possible, same reagents, protocols, personnel, technologies etc. be used for all the batches in a study. In addition, making sure that biological variables (example disease status) do not overlap with technical variables (example sequencing date) will assure that the results obtained are due to biological differences between samples and not technical differences. As biological and computational methods become more efficient, we will be able to draw more robust conclusions from analysis of complex metagenomic communities, but issues relating to sampling and sequencing procedures and the choice of the methods used for bioinformatic and statistical analysis of the community in question should be considered, not only in the beginning but also throughout the course of any metagenomic study.

Advances in sequencing technology and its effect on analysis of complex ecosystems: Initially environmental gene sequencing focused on specific genes (often the 16S rRNA gene) to obtain a profile of the microbial diversity in the environmental sample. More recently, however, “shotgun Sanger sequencing, massively parallel pyrosequencing”, or

Illumina sequencing [17] have been used to obtain sequences of all genes from all members of sampled communities. These studies, whether the focus is on a single gene or on all the genes, revealed that culture based methods missed a majority of the microbial diversity within the environment[18].

Shotgun sequencing, the approach which had been used to sequence many cultured microorganisms[19]as well as the human genome [20], randomly shears the extracted genomic DNA into many short sequences before sequencing them. These short fragments were sequenced by Sanger sequencing [21] in earlier studies but in the recent past high-throughput sequencing methods are being increasingly used [13], [22]. The Sanger sequencing method (Sanger et al., 1977) is based on synthesizing DNA based on a single stranded template while randomly incorporating chain terminators, and the different fragment sizes generated by this sequencing method coincide with to the chain-terminator locations. In the last decade, the average length of a sequencing-read generated by Sanger sequencing has increased from around 450bp to 850bp. Due to the fact that the Sanger method runs one sequencing reaction at a time, large metagenomic studies that utilize Sanger sequencing could only be carried out at large genome centers with hundreds of sequencing machines, all of them working simultaneously to sequence the metagenome. Until now, the largest such metagenomic study to have utilized Sanger sequencing is the Sorcerer II Global Ocean Sampling (GOS) expedition [23], lead by Dr Craig Venter (well known for his role in the Human Genome Project). The enormous size of this study can be appreciated by the fact that just the pilot project of this study (conducted in the Sargasso Sea) yielded DNA from about 2000 different species, 148 of which were completely novel bacteria[24]. This study ended up increasing the size of protein

databases to almost twice their original size by adding millions of predicted protein sequences and thousands of protein families to the protein databases.

New sequencing approaches, made possible by parallel advances in fields of enzymology, imaging and microfluidics, have increased sequencing capacity but are not associated with the huge infrastructure involved in earlier sequencing methods. Most sequencing processes involve an initial amplification step that amplifies the DNA. In the Sanger method, this is usually done by cloning, where the DNA is incorporated into a plasmid and the clones are then grown. Due to a number of reasons (fragment toxicity, replication inhibition etc.) the bacteria, mostly E. coli, into which the plasmids are transformed, can selectively amplify certain fragments of DNA inducing a bias in this step. To overcome the aforementioned shortcomings of the in-vivo methods, Margulies et al developed a high throughput strategy for in-vitro amplification that has an added advantage of also being inexpensive relative to Sanger sequencing. This method [25] is commonly known as 454 pyrosequencing after 454 Life Sciences (Branford, CT, USA), the company that commercialized this technology. With the high accuracy, low cost, and relatively long reads associated with some “next generation” methods like 454 sequencing and Illumina sequencing, many researchers have migrated away from traditional Sanger capillary sequencing instruments and toward these sequencing platforms for a variety of their genome projects. Forest Rohwer’s group at San Diego State University were the first to use next generation sequencing, pyrosequencing developed by 454 Life Sciences[25], for sequencing community DNA[26]. Even though the 454 sequencing method generates shorter sequence lengths, it compensates for that by generating very large number of sequences compared to traditional Sanger sequencing

methods. The newly available titanium platform, from 454, allows reads as long as 400bp and is therefore beginning to approach the read lengths reachable through traditional Sanger methods[27]. More recently the Illumina sequencing technology[17] is being increasingly used for shotgun metagenomic studies[22] including for the Human Microbiome Project; due to its lower cost and lower error rates than the pyrosequencing method.

Bioinformatics methods and challenges in community analysis: Analysis pipelines of many early metagenomics studies concentrated on gathering enough sequence information to characterize complete genomes from the concoction of metagenomic sequences. This was possible for low complexity environments, such as an acid mine drainage ecosystem [28], by using various complicated “binning” methods (grouping sequences based on oligonucleotide signatures). Whereas in more complex environments like soil or ocean samples, assembly still remains one of the major analysis limitations. Sequence data from complex environments, due to high levels of microbial diversity, is heterogeneous and in most cases contains an unequal representation of the constituent species. In addition, organisms in a complex environment frequently belong to closely related strains, whose genomes are highly similar, making it practically impossible to construct assemblies of each organism present in a sample. Also, viruses and/or inserted phages, if present, increase the possibility of generating chimeric contigs[29] that further impede assembly. The short-reads associated with newer generation sequencing methods, like 454 sequencing and Illumina sequencing, impose further complications. Due to the limitations in assembly of metagenomic data, gene prediction methods used in metagenomic analysis have been adapted to work with large numbers of fragmented

genes on short sequences. However, due to the phylogenetic diversity in samples it is difficult to find appropriate training sets for “intrinsic” gene finding in metagenomes. Consequently, extrinsic gene finding strategies that find coding regions based on their similarity to genes and coding regions in a reference database have been used. Some studies (e.g.[30], [31]) skip gene prediction altogether and focus only on the ‘known fraction’ of their dataset by limiting the downstream analyses to the BLAST annotated portion of their reads. These studies rely on direct classification of raw reads by homology to existing sequences in sequence databases[32] but the disadvantage of this approach is that it will miss genes from novel organisms that have no close relative (homologs) in the sequence databases.

Taxonomic profiling of metagenomic reads: Assessing the composition of the community in question is one of the crucial steps in understanding the role that microbes play in their environment. Traditionally, 16S rRNA gene sequences have been used for taxonomic assignment in genomes extracted from cultured organisms [33]. The sequencing of 16S rRNA genes from new species is made possible by the presence of highly conserved regions at several positions, well-located, along the gene [34]. The conservation of these regions allows one to design and use broadly targeted oligonucleotide primers that work on a wide diversity of species for both sequencing and amplification by the polymerase chain reaction (PCR). The amplified products can then be characterized in multiple ways; such as through restriction digestion[35], denaturing gradient gel electrophoresis[36-37], hybridization to arrays[38], or sequencing [39],[40],[41],[42]. As sequencing continues to decrease in cost and difficulty, it has

become the preferred option and therefore we focus only on sequence analysis in this dissertation.

The length of the gene targeted using 16S rRNA gene sequencing, not surprisingly, has been dependent on the sequence length options offered by the sequencing technology available at the time the study was initiated. This is corroborated by the fact that earlier sequencing studies, targeting the 16S rRNA gene, captured either the entire or most of 16S rRNA gene, using the longer read-length associated with traditional Sanger sequencing. Recently, with the rapid development of next generation sequencing technologies, uncultured bacteria from complex environments have been sequenced at a much lower cost than Sanger dideoxy sequencing. One of the earliest examples of the use of pyrosequencing in surveying microbial diversity is the exploration of the "deep sea" by Sogin and colleagues [43]. One of their reasons for choosing the V6 region for the study is that the shorter length of V6 variable region of the 16S rRNA (~65bp), compared to the other 16S variable regions, makes it amenable for capture by the 100-bp reads generated by the pyrosequencing technology (GS-20), available at that time. More recently, the read length of 454 pyrosequencing machines has been increased to an average of 250bp (GS-FLX) and later to 400bp (454-titanium). This opened up more options for primer design and allowed the possibility of targeting regions of the 16S rRNA gene other than just the V6 region[44]. Using these newly available technologies, a vast numbers of "partial sequences" from 16S rRNA genes of environmental DNA have been generated and analyzed. The use of partial 16S rRNA sequences has been feasible due to studies that found that even fragments of the 16S rRNA gene can be used as substitutes for the full-length sequence, in many community analyses [45-46]. The

pyrosequencing approach has been used to target a wide range of microbial communities and variable regions of the 16S rRNA gene, including the V6 region in deep-sea vents microbial communities [43]; V1, V2, V6 and V3 regions in human gastrointestinal tract [39],[47],[48]as well as the V9 region in soil-derived microbial DNA[49].

Whole genome sequence based methods that utilize the random or shotgun sequences, generated from the entire DNA of the environmental sample[10], for characterization of the community, have been suggested as a potential alternative for rRNA gene sequence-based studies. These methods, also known as “metagenomic methods”, are indeed very powerful in that they bypass some of the limitations of PCR methods and, in the process, generate sequence data of many genes, including the 16S rRNA gene, from the many organisms present in a community. Taxonomic profiling of a community using random whole genome sequence reads can not only characterize “Who is there?” but can also be used to predict “What they are doing?”[50]. In some cases, application of shotgun metagenomics has led to the discovery of novel lineages of organisms that have been entirely gone undetected by rRNA gene PCR methods [51].

Metagenomics is most likely to help us reveal the complex microbial communities, inhabiting nearly every environment and organism on Earth, that have been invisible so far due to the limitations of pre-existing technologies. Extracting all the possible information from metagenomic libraries will continue to be difficult, mainly because of the massive size and complexity of the datasets. Greater sequencing depth enabled by the lower cost and higher resolution of new technologies would make it possible to detect the rare yet important members of our biosphere. But more importantly, improvements in bioinformatics tools will make it easier to interpret the metagenome sequence data and in

some cases may help assemble whole genomes from metagenomic sequence data. Even in communities where assembly is not possible bioinformatic tools, by unearthing the microbial composition of the community in question, can help us move closer towards a better understanding of the role microbes play in an environment.

In Chapters#1 and #3 of this dissertation we discuss metagenomic analyses of some understudied microbial communities, during the course of which we touch upon some of the bioinformatic challenges, mentioned above, which arise during these analyses. In Chapter 2 we talk about batch-effects that are one of the major challenges faced during metagenomic analysis and how such effects can mask the true biological signal. In Chapter#4 we provide a comparative exploration of some the taxonomic composition estimating tools used during metagenomic analyses to exemplify the effect of analysis choices on the results of a metagenomic study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BMI	body mass index
DGGE	Denaturing Gradient Gel Electrophoresis
RISA	Ribosomal Intergenic Spacer Analysis
FISH	Fluorescent in situ hybridization
TRFLP	Terminal restriction fragment length polymorphism
RNA	Ribonucleic acid
UNC	University of North Carolina
DHS	Diet and Health Study
IRB	Institutional Review Board
DNA	Deoxy-Ribonucleic Acid
CAGE	Core for Applied Genomics and Ecology
PCR	Polymerase Chain Reaction
QC	Quality Control
RDP	Ribosomal Database Project
OTU	Operational Taxonomic Unit
BLAST	Basic Local Alignment Search Tool
EPA	Evolutionary Placement Algorithm
q-PCR	Quantitative Polymerase Chain Reaction
PCoA	Principle Coordinate Analysis
FDR	False Discovery Rate
WHR	Waist to hip Ratio

SEM	Standard Error of the Mean
SD	standard deviation
PCA	Principle Component Analysis
NPDES	National Pollutant Discharge Elimination System
CBOD5	Carbonaceous Biochemical Oxygen Demand (5 day test)
CFU	Colony Forming Units
COD	Chemical Oxygen Demand
NCBI	National Center for Biotechnology Information
GOS	Global Ocean Survey
EPBR	enhanced biological phosphate removal
BNR	Biological Nutrient Removal
MEGAN	MEta Genome ANalyzer
CARMA	Characterizing short Read Metagenomes
LCA	Least Common Ancestor
pHMM	Profile Hidden Markov Model
EGT	Environmental Gene Tags
Pfam	Protein Family
DZ	Dizygotic
MB	Megabases
WGS	Whole Genome Sequences

CHAPTER 1: INCREASED MICROBIAL RICHNESS IS ASSOCIATED WITH HUMAN COLORECTAL ADENOMAS

1.1 Abstract

Differences in gut microbial community composition have been linked to many important human diseases including obesity, Crohn's disease, Ulcerative Colitis [52], [53], [54] and colorectal cancer. Previous studies that suspected a link between commensal gut bacteria and colorectal cancer, however used low throughput methods [55], [56], [57]. In this study, we employed 454 titanium pyrosequencing of the V1-V2 region of the 16S rRNA gene to characterize adherent bacterial communities from mucosal biopsies of 33 adenoma subjects and 38 non-adenoma subjects. We found 87 taxa (including known pathogens) that had significantly higher relative abundances in cases vs. controls while only 5 taxa that were more abundant in control samples. In addition, adenoma samples had a pronounced increase in average microbial richness suggesting that conditions associated with colorectal adenomas create an environment in which potentially pathogenic microbes can flourish. Intriguingly, the magnitude of the differences between adenoma case and control in the gut microbiota was more pronounced than differences in the microbiota associated with patient obesity. Because the microbial signature associated with colorectal adenomas is generally distinct from microbial signatures associated with known risk factors such as increased body mass

index (BMI), these results suggest that next-generation sequencing of the gut microbiota has potential utility as a diagnostic tool indicating the presence of adenomas.

1.2 Background and significance

The human microbiome, the microbes that are associated with the human body, outnumber our own "human" cells 10 to 1[58] and provide us with a wide array of vital metabolic functions that we are lacking in[12]. The role that these "beneficial" microbes, play in health and disease, has been explored in the past, but only recently has the technology reached a point where the species present within an individual's microbiome can not only be accessed but identified [59], [12], [60], [61], [62], [63]. Recent research has shown that the relationship between the gut bacteria and humans is not just commensal (non-harmful coexistence), but is in fact symbiotic (mutually beneficial)[64]. For instance, microbes living in the our gut help us in digestion of food, in disruption of toxic compounds and in combating disease-causing pathogens[65]. Changes in these microbial communities may be responsible for digestive disorders [66-67],[68], skin diseases [63], obesity [69],[8],[59],[7],[70],[71],[72] and a range of "immuno-pathologic" conditions including inflammatory bowel diseases [73],[74],[75-76]. These studies suggest that each individual person is a "microbial island", meaning that each has their own unique bacterial signature just as each individual has a unique fingerprint. However, our gut microbiomes share a core group of genes that carry out some core functions and the differences in this "core set" can define different physiological states or phenotypes (for example lean and obese)[52]. In spite of the strong individual differences in the microbial community, researchers studying the human microbiome often perform cross-sectional, "case-control" studies, which look for differences in bacterial populations

between patients who have a specific disease and those who do not[68], [77-78]. These studies have shown distinct microbial signatures of disease groups that separate cases from controls in diseases such as periodontal disease and gastric cancer [79], [80], [67]. The results from these studies indicate that disruption of the human microbiome levels plays a crucial role in human health and disease and that these changes can be indicators of the disease status in the human hosts. As a possible mechanism, Mazmanian et. al. have proposed that "the equilibrium between potentially harmful and potentially beneficial bacteria in the gut mediates health versus disease"[81]. Under this model, if the balance is altered by changes, for instance due to diet, stress or antibiotics, then the immune response in the intestines is also changed leading to inflammation. This change in host-microbe relationship, called "dysbiosis", has been associated with numerous gastro-intestinal diseases like inflammatory bowel disease [77] colon cancer [55] obesity [7-8, 70] and diabetes [78]. Chronic inflammation leads to cancer, and this mechanism has been suggested as a possible trigger for inflammation and colon cancer in animal models [82].

Colorectal cancer is the second most common cancer in women and third most common cancer in men in the Europe and is the second leading cause of death resulting from cancer in both sexes[83], in developed countries. Although age, tobacco and alcohol consumption, physical activity and body weight are considered important risk factors for colorectal cancer[84], the most significant risk factor happens to be diet [85], [86] . In addition to the various factors mentioned above, the role of host associated microbiota has also been frequently proposed as a critical factor in colorectal cancer [55],[87],[88],[57]. Recent studies have investigated the possible role of the microbial

component of the colon in Colorectal Cancer [55] and have used culture independent approaches to explore the distal gut's microbiome diversity and stability in individuals with colorectal cancer [88]. These studies[87], [88], used 16S rRNA gene denaturing gradient gel electrophoresis (DGGE) and ribosomal intergenic spacer analysis (RISA) to explore of the microbial diversity in the fecal samples in case and control subjects. Recent research on the mucosal adherent microbial component of the colon [89] showed that the bacterial community profiles of healthy individuals are stable along the length of the colon. While each individual has a distinct bacterial profile, there is some overlap between the mucosal-associated bacterial communities among individuals [87].

In a recently published study[57], our collaborator Dr Keku Temitope and her colleagues characterized the adherent bacteria in normal colon and in the diseased colon by fluorescent in-situ hybridization (FISH) analysis of the 16S rRNA genes as well as by terminal restriction fragment length polymorphism (TRFLP) and Sanger sequencing of 16S rRNA clones. Their study showed that a distinct microbial signature is associated with colorectal adenomas. The work described in this Chapter, is a further extension of Dr. Temitope's study via the utilization of second-generation sequencing technology, to provide deeper coverage of bacterial communities and to characterize the gut microbial communities of a larger set of patients.

1.3 Materials and Methods

1.3.1 Patient characteristics

Subjects were screening colonoscopy patients at UNC Hospitals who agreed to participate in the Diet and Health Study (DHS V) and the characteristics of these subjects are shown in Table 1.1. The enrollment procedure as well as colonoscopy and biopsy

procedures and sample collection have been previously described [90], [57]. The study was approved by the Institutional Review Board (IRB) at the University of North Carolina, School of Medicine (Protocol #05-3138).

1.3.2 DNA extraction and sequencing

Bacterial genomic DNA was extracted from mucosal biopsies; the biopsies ranged in weight between 10-20 mg. Two biopsies per subject were used for bacterial DNA extraction and these were placed in lysozyme (30mg/ml; Sigma, St. Louis MO) for 30 minutes. The biopsy-lysozyme mixture was homogenized on a bead beater (Biospec Products Inc., Bartlesville, OK) at 4,800 rpm for 3 minutes at room temperature followed by DNA extraction using the Qiagen DNA isolation kit (cat # 14123) per the manufacturer's recommended protocol. The mucosal adherent microbiome was analyzed by Roche 454 titanium pyrosequencing of 16S rRNA tags from genomic DNAs. Pyrosequencing [25] was conducted at the University of Nebraska Lincoln Core for Applied Genomics and Ecology (CAGE). We amplified the V1-V2 region (F8-R357) of the 16S rRNA gene from mucosal biopsies followed by titanium-based pyrosequence analyses. The 16S primers contained the Roche 454 Life Science's A or B Titanium sequencing adapter (italicized), followed immediately by a unique 8-base barcode sequence (BBBBBBBB) and finally the 5' end of primer A-8FM, 5' - CCATCTCATCCCTGCGTGTCTCGACTCAGBBBBBBBBAGAGTTTGATCMTGGC TCAG-3' and B-357R, 5'- CCTATCCCCTGTGTGCCTTGGCAGTCTCAGBBBBBBBBCTGCTGCCTYCCGTA-3'. Each DNA sample was amplified with uniquely barcoded primers, which allowed us to mix PCR products from many samples in a single run.

1.3.3 Data filtering

1.3.3.1 Sample filtering

We screened all the samples for a batch-effect that correlated with the date of submission to the sequencing center. Samples were shipped on 3 separate dates from Chapel Hill to the sequencing center in Nebraska. Samples shipped on one particular date (09/30/2009) were found to cluster separately from samples shipped on other dates (06/10/2008 and 7/21/2008). The DNA stocks of these 2 groups of samples were also stored in different freezers at the Chapel Hill lab. In addition, the sum of Bacteroidetes and Firmicutes observed in samples shipped on this date was much lower than we would expect based on both previously published human gut microbial 454 datasets and our own 454 datasets. Sequences generated from samples sent to the sequencing center on this date were therefore removed from further analysis. Leek et al. recently showed the importance of screening high throughput datasets for batch-effects [91] and screening for batch-effects indeed proved useful in removing the technical artifacts from our dataset. The descriptive characteristics and of the 71 samples, 33 cases and 38 controls selected after sample filtering, are shown in Table 1.1.

1.3.3.2 Sequence filtering

1.3.3.2.1 RDP Pipeline

The first step in the data analysis process involved a preliminary QC (quality control) filter (downstream of the Roche-454 GS-FLX software filtering). We removed sequences from our dataset if there were any Ns in the sequence or the 5' primer did not exactly match the expected 5' primer or if the average quality score was less than 20. We then removed the 5' primer sequence from our reads that have survived above filtering. Only

trimmed filtered sequences with a length between 200-500bp were kept in our data set for RDP analysis.

1.3.3.2.2 OTU Pipeline

We removed sequences from inclusion in the OTU dataset if there were any Ns in the trimmed sequence or if the 5' primer did not exactly match the expected 5' primer. As recommended by Kunin et. al.[92], sequences were end-trimmed with the Lucy algorithm [93] at a threshold of 0.002 (quality score of 27). Only reads with trimmed lengths between 150 and 450 were retained for OTU analysis. Table 1.2 shows the number of sequences removed by our RDP and OTU pipelines.

1.3.4 Bacterial Identification

The sequences in our dataset were given taxonomic assignments based on two methods.

1.3.4.1 RDP assignment method

Sequences that have been filtered using the RDP pipeline (Table 1.2) were submitted to the RDP Classifier 2.1 algorithm for taxonomic identification at various taxonomic levels. Sequences assigned in each sample to various taxa, from phylum level and genus level, were counted at the RDP confidence threshold of 80.

1.3.4.2 OTU assignment method

OTU analysis is more sensitive to sequencing error[92] and we therefore applied additional QC steps in our OTU analysis pipeline (Table 1.2). Sequences filtered through the OTU pipeline were submitted to Abundant OTU (http://omics.informatics.indiana.edu/AbundantOTU/) for assignment of each sequence to operational taxonomic units (OTUs; 97% identity). Sequences assigned in each sample to

various OTUs were counted and then normalized and log transformed (see Data Preprocessing), before proceeding to further downstream analyses. Consensus sequences generated by AbundantOTU during construction of OTUs were submitted to RDP classifier 2.1 to assign taxonomy to each of the OTU groups. Consensus sequences of the 613 OTUs generated by AbundantOTU (available as Sanapareddy_SupplementaryDataFile1) were also submitted to ChimeraSlayer [94] (http://microbiomeutil.sourceforge.net/) and the 9 consensus OTUs identified by chimera slayer as chimeras were removed from our dataset. In addition consensus sequences of 4 OTUs on BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) search against the Silva reference 16S database failed to match with >97% sequence identity so these were also removed from further analysis. This left a total of 600 OTUs.

1.3.5 Richness and Evenness

Shannon-Wiener Diversity Index, H, was calculated using the equation, $H = -\sum Pi\ (lnPi)$, where Pi is the proportion of each species (taxa) in the sample. Richness was calculated as the number of OTUs, genera or phyla observed in 2,636 sequences (where 2,636 is the number of sequences seen in the sample with the fewest sequences). For each sample, 2,636 sequences were randomly chosen 1,000 times and the average number of OTUs, genera or phyla observed over these 1,000 permutations was reported as richness.

Evenness measures how evenly the individuals are distributed among the different species/taxa and is calculated by $J= H'/Log\ (S)$ where H' is Shannon diversity and S is the number of species or taxa in each sample. Wilcoxon-tests and Student's t-tests were performed to compare the mean similarities of the groups, case and control. The false

discovery rate was set at 10% using the Benjamini and Hochberg procedure[95] to avoid type 1 error due to multiple comparisons on a single data set.

1.3.6 Data Preprocessing

1.3.6.1 Normalization

Raw counts were normalized then log transformed using the normalization scheme mentioned below, before proceeding with the rest of the analyses.

LOG10 ((Raw count / # of sequences in that sample)*Average # of sequences per sample +1).

1.3.6.2 Removal of rare taxa

In order to minimize the number of null hypotheses for which we would need to correct for multiple hypothesis testing, we removed rarely occurring taxa that occurred in so few patients that they could not be significantly associated with case-control or obesity phenotypes. In all of our analyses (except richness calculations), we therefore only included taxa which occurred at least once in 25% of all samples. For the RDP approach, 9 phyla and 100 genera met this criterion. For the OTU approach, 371 OTUs met this criterion.

1.3.7 Tree Generation

For each of the 371 consensus sequences from OTUs that met the above criteria, BLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to find the top 10 hits in the Silva reference tree release 104 (http://www.arb-silva.de/download/arb-files/). In this way, we identified a set of 3,594 aligned sequences to serve as our reference tree. The program align.seqs within MOTHUR (http://www.mothur.org/) was used to align the 371 AbundantOTU consensus sequences that passed all QC steps, to these 3,594 aligned

sequences as extracted from the Silva reference alignment. With custom Java code based on the Archaeopteryx code base (http://www.phylosoft.org/archaeopteryx/), we removed all but the 3,594 sequences from the Silva reference tree. We then uploaded the alignment of the 3,594 reference sequences plus the 371 AbundantOTU sequences to the RaxXML EPA server (http://i12k-exelixis3.informatik.tu-muenchen.de/raxml), which uses maximum likelihood to place new sequences within a reference tree. Custom Java code (available upon request) was used to add RDP calls from each consensus sequence (Appendix A, Supplementary Figure 5) and coloring by false discovery rate (Figure 1.2, Appendix A, Supplementary Figure 5) to the tree. Trees were visualized with Archaeopteryx. Leaf nodes in Supplementary Figure 5 (Appendix A, Supplementary figures) are labeled with the RDP call of the consensus sequence at 80% confidence.

1.3.8 UniFrac Analysis

The tree generated from the 371 OTU consensus sequences (using RaxXML EPA server described above) along with the environment file with the abundance information of each of the 371 OTUs within the case and control environments were submitted to UniFrac [96] and Fast UniFrac to see if cases cluster separately from controls. We ran 100 permutations on the abundance weighted tree using the UniFrac significance test.

1.3.9 Data Validation

1.3.9.1 Real-time quantitative PCR validation

q-PCR primers were designed based on no less than 95% sequence similarity from bacterial 16S ribosomal DNA sequence alignments obtained from pyrosequencing. To measure the abundance of a specific taxon, three primer pairs where designed: one generic for all bacterial groups (Universal Primer): [EUB341-F 5’-

CCTACGGGAGGCAGCAG-3' EUB518-R 5'-ATTACCGCGGCTGCTGG-3'] and three taxon-specific primer pairs: first for the Helicobacter genus (Heli_F 5' AGTGGCGCACGGGTGAGTA 3'

Heli_R 5' GTGTCCGTTCACCCTCTCA 3'), the next one for the Acidovorax genus (Aci_F 5'-TGCTGACGAGTGGCGAAC-3' Aci_R 5'-GTGGCTGGTCGTCCTCTC-3') and another for the Cloacibacterium genus (Clo_F 5'-TGCGGAACACGTGTGCAA-3' Clo_R 5'-CCGTTACCTCACCAACTAGC-3').

10 µL PCR reactions were prepared containing 100ng of DNA extracted from colonic mucosal biopsies, 10 µM of each primer, and 5 µL of Fast-SYBR Green Master Mix (Applied Biosystems). Cycling conditions were: 1 cycle at 95°C for 10 minutes followed by 45 cycles of 95°C for 15 seconds, 60°C for 1 minute, and 72°C for 30 seconds. A single dissociation curve cycle was run as follows: 95°C for 30 seconds, 60°C for 30 minute, and 90°C for 30 seconds. A pool of samples was prepared to serve as the standard for the qPCR by mixing equal volumes from each sample. Abundance of a specific taxon was calculated by the delta-delta threshold cycle (ΔΔCt) method[97] in which: ΔΔCt = (CtTSE – CtUE) – (CtTSP – CtUP) .Where: CtTSE: Ct of experimental samples for taxon-specific primers, CtUE: Ct of experimental samples for universal primer, CtTSP: Ct for DNA Pool for taxon-specific primers, CtUP: Ct for DNA pool for universal primers. Theoretically, the abundance of a taxon is 2−ddCt.

1.3.10 Nucleotide sequence accession numbers

All 454 pyrosequences from this study are available in the Genbank database under the accession # SRS 166138.1-172960.2.

1.3.11 Statistical analyses

The diversity indices, richness and evenness, were calculated using JAVA implementations (available upon request). Kruskal-Wallis, Wilcoxon and Student's t-tests were performed using JMP 8.0 (SAS Institute, Cary NC) to compare the mean similarities of the groups, case and control. Regression and correlation analyses were performed using JMP 8.0 (SAS Institute, Cary NC) and in R (Open Sourced Statistical software).

1.4 Results

To evaluate associations between the gut microbiota and the presence of adenomas, we collected mucosal biopsies from the same region (~10-12 cm regions from the anal verge) from 33 adenoma subjects and 38 controls. Our initial analyses looked at global signatures of the entire microbial community. At the phylum, genus and OTU levels we found significant differences in richness (i.e. the number of taxa present in a sample), but no differences in evenness (i.e. how evenly distributed taxa are within a sample), between cases and controls (Figure 1.1; Appendix A, Supplementary Figures 1 & 2). In order to see whether case samples cluster separately from control samples, we used UniFrac[96] to cluster our sequences based on their placement in the phylogenetic tree shown in Figure 1.2. Running 100 permutations on the abundance weighted tree using the UniFrac significance test resulted in a p-Value of 0.02 suggesting a marginally significant separation between cases and controls when considering all of the nodes of the phylogenetic tree. Similarly, weak clustering was seen when we used principle co-ordinate analysis (PCoA) on the same tree using FastUnifrac (Appendix A, Supplementary Figure 3).

We next asked which individual bacterial taxa were different between cases and controls. By examining the results of the RDP classification algorithm [46] at the phylum level, we observed at a 10% false discovery rate threshold that cases had higher relative abundance of TM7, Cyanobacteria and Verrucomicrobia compared to controls (Appendix A, Supplementary Table 1). At the genus level at a 10% false discovery rate threshold, the relative abundance levels of 30 genera including Acidovorax, Aquabacterium, Cloacibacterium, Helicobacter, Lactococcus, Lactobacillus and Pseudomonas were higher in cases vs. controls (Appendix A, Supplementary Table 2). Remarkably, only one genus, Streptococcus, had a higher relative abundance in the control group. In order to validate these pyrosequencing results, we developed qPCR assays for a subset of observed genera that were significantly different in their relative abundances between cases and controls (i.e., Helicobacter spp, Acidovorax spp and Cloacibacteria spp.). We observed the expected correlations between the two methods (Appendix A, Supplementary Figure 4), validating the results of our pyrosequencing approach.

We also performed an analysis of Operational Taxonomic Units (OTUs), which are clusters of sequences in which the average percent identity of all of the sequences within a cluster is >=97%. Our analysis at the OTU level at a 10% false discovery rate threshold found 87 OTUs with significantly higher relative abundance in cases vs. controls and only 5 OTUs higher in controls (Appendix A, Supplementary Table 3). When we used the RDP classification algorithm to classify the consensus sequence for each of the 92 significantly different OTUs, bacteria with higher relative abundance in cases were mostly members of the phyla Firmicutes (42.6%), Bacteroidetes (25.5%) and

Proteobacteria (24.5%) (Figure 1.2, Appendix A, Supplementary Figure 5). A rank-abundance curve demonstrates that the OTU differences between cases and controls (significant at 10% FDR) are entirely in low abundance taxa (Appendix A, Supplementary Figure 6). This observation explains why there are differences between case and control in richness (Figure 1.1), which depends on the total number of taxa observed, but not evenness, which is more sensitive to changes in high-abundance taxa.

Since obesity is a risk-factor for development of colorectal cancer, and changes in the human microbiome have been associated with obesity [52], [98] we evaluated the relationship between the relative abundance levels of the individual taxa and the risk factors, BMI and Waist-to-Hip Ratio (WHR). We classified subjects into one of three BMI categories; Normal (BMI<25), Overweight (BMI = 25-29) and Obese (BMI 30 and above) and three WHR levels; low, medium and high based on accepted thresholds (http://www.bmi-calculator.net/waist-to-hip-ratio-calculator/waist-to-hip-ratio-chart.php). For each OTU, the non-parametric Kruskal-Wallis test was performed between the three groups for BMI and WHR. There were no OTUs that showed significant differences between the various BMI and WHR risk factor categories even if we were to set a false discovery rate threshold as high as <200% (Appendix A, Appendix A, Supplementary Tables 4 & 5). Likewise, there were no significant differences in the diversity measures, richness and evenness, between the various risk factor categories (Figures 1.3 & 1.4). Finally, regressions between BMI values and WHR values against each taxa at the OTU level also showed no significant association between the OTUs with either BMI or WHR at an FDR threshold of <10% (Appendix A, Supplementary Figures 7 & 8, Appendix A, Supplementary Tables 6 & 7).

1.5 Discussion

Taken together, these findings demonstrate that the development of adenomas is associated with changes in the relative abundance of various taxa, including pathogens, present in the gut mucosa and that these changes are distinct from those associated with obesity. Analogous to the mechanism suggested for inflammatory bowel diseases[99], a potential explanation for this observation could be that the presence of adenomas compromises gut mucosal immunity, leading to an increased relative abundance in known pathogens such as Pseudomonas, Helicobacter, Acinetobacter (Appendix A: Appendix A, Supplementary Table 2, Supplementary Table 3) and other genera belonging to the phylum Proteobacteria (Figure 1.2). Alternatively, the presence of these pathogens may directly increase the risk of adenoma development by changing the gut environment. For example, Helicobacter has a much higher relative abundance in cases vs. controls (Appendix A, Supplementary Tables 2& 3) consistent with previous studies, which implicate the role of this bacterium in colorectal adenomas[100],[101],[102]; a possible explanation for this association is that this microbe alters the pH of the gastrointestinal tract[103],[104]. Acidovorax spp, another member of the bacterial signature identified as significantly different between case and control in this study, is a flagellated, Gram-negative acid-degrading member of the phylum Proteobacteria. Although, not much is known about its clinical epidemiology and pathogenicity in humans, it has been associated with induction of local inflammation [105], [106]. Lactobacillus, another taxa that we found to be higher in case than control, is an acid producing bacteria known to lower gut pH and regulate the growth of other bacteria. While Lactobacillus is generally considered a beneficial microbe, [107], [108] its

presence in this case may help to lower pH to create favorable conditions for bacterial dysbiosis. This is consistent with suggestions by Duncan and co-workers [109] that bacteria that grow in acidic pH create an environment that can be exploited by more low pH-tolerant microbes.

While further experiments will be required to determine if and how increased microbial richness causes the development of adenomas, our observation that the microbial signature associated with adenomas is largely distinct from that associated with obesity suggests that next-generation sequencing of microbial communities may have considerable value as a diagnostic that can separate risk-factors from the actual presence of adenomas.

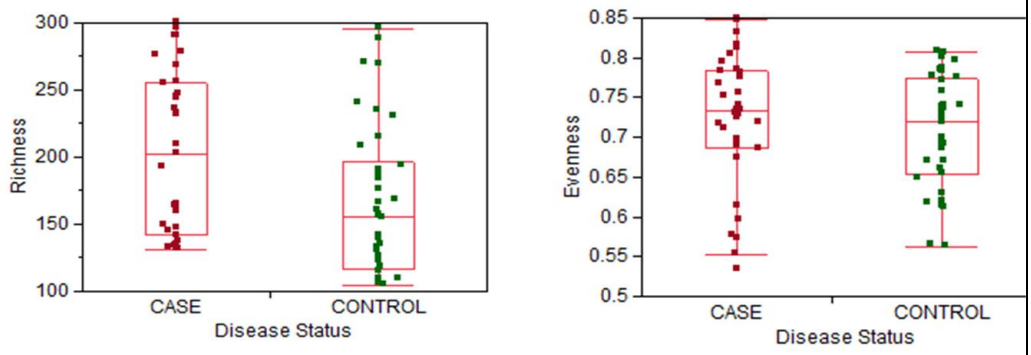


FIGURE 1.1: Richness (left panel) and evenness (right panel) for the OTUs observed in our study for cases (n=33) vs. controls (n=38). The x-axis is proportional to the number of subjects in each category. By the Wilcoxon test, cases had a significantly higher richness (p= 0.0061) than controls, but there was no significant difference in evenness (p = 0.36).

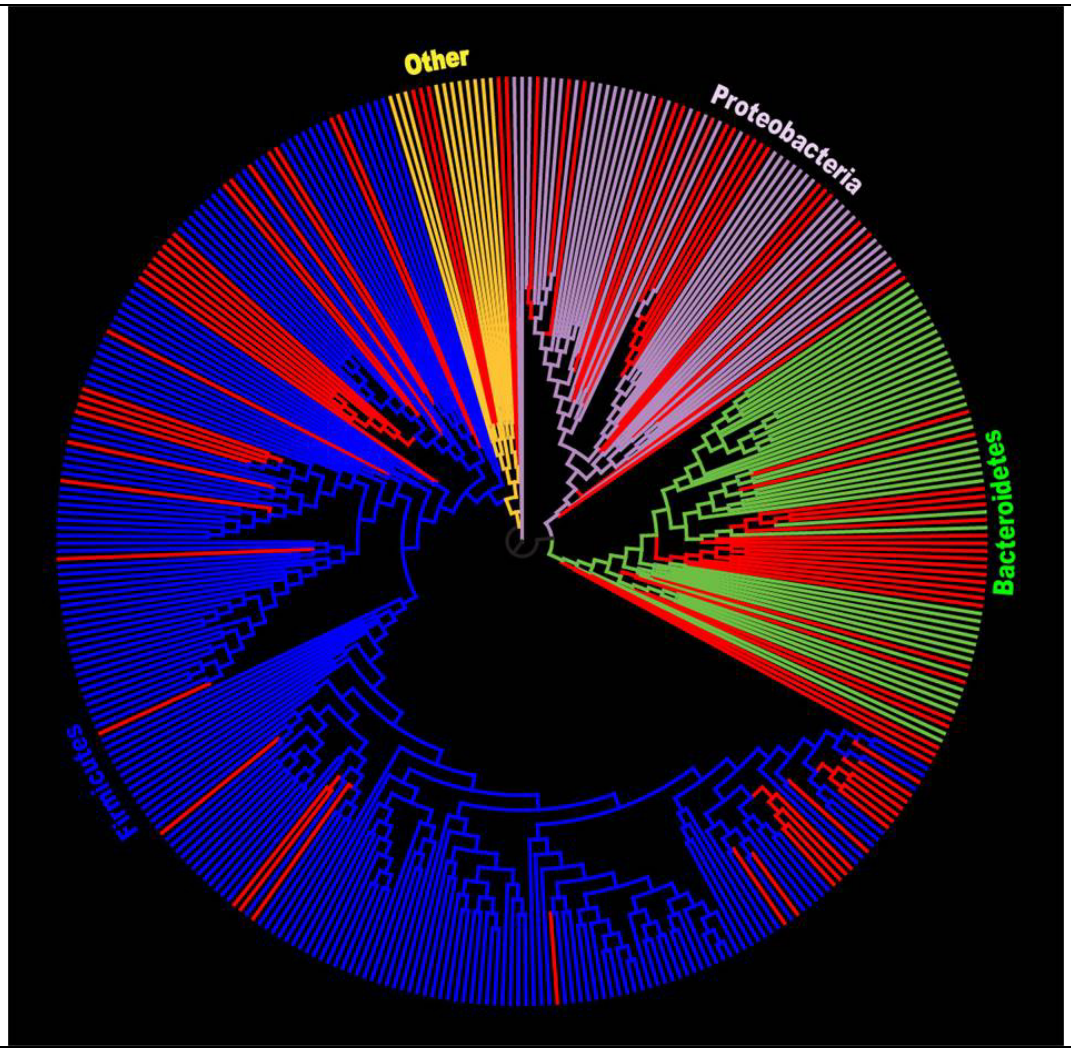


FIGURE 1.2: Maximum likelihood tree generated from the 371 OTUs (OTUs that were observed in at least 25% of our patients). The tree was generated using the RaxXML EPA server (http://i12k-exelixis3.informatik.tu-muenchen.de/raxml) (see methods). Branches are colored based on RDP Phylum level assignments. Red colored branches represent OTUs significantly different between cases and controls within each Phylum (at 10% FDR).

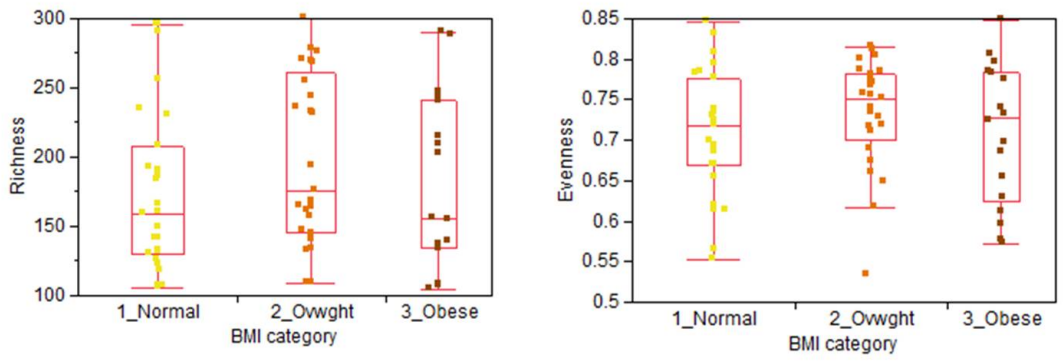


FIGURE 1.3: Richness (left panel) and evenness (right panel) at the OTU level, in Normal (n=27) vs. Overweight (n=25) vs. Obese (n=18) BMI categories. No significant difference was seen by the Kruskal-Wallis test in richness (p = 0.21) or evenness (p = 0.42) between the 3 categories.

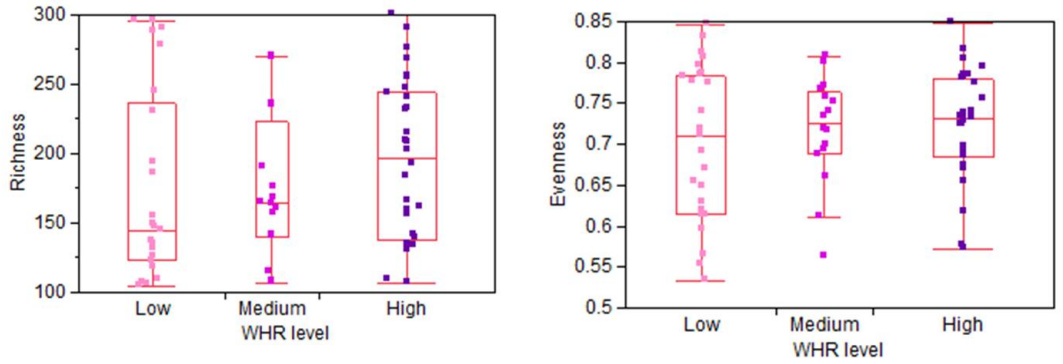


FIGURE 1.4: Richness (left panel) and evenness (right panel) at the OTU level, in Low-Risk (n=25) vs. Medium-Risk (n=16) vs. High-Risk (n=30) Waist-to-hip ratio categories. No significant difference was seen by the Kruskal-Wallis test in richness (p = 0.26) or evenness (p = 0.76) between the 3 categories.

TABLE 1.1: Descriptive characteristics of the study participants, cases (33) and controls (38). p-Values are based on t-tests between case and control (age, WHR and caloric intake) or the Chi square test (% Male and %BMI). The *p-Value for BMI is from the chi-quare test comparing across the groups. Caloric intake is reported as kilocalories (kcal) and is based on responses from a food frequency questionnaire [110] that was administered to subjects during phone interviews.

Characteristics	Case (n=33)	Control (n=38)	p-Value*
Age (mean, SEM)	57.45 (1.11)	55.70 (1.08)	0.26
Male (%)	60.61	50	0.54
WHR (mean, SEM)	0.94 (0.01)	0.90 (0.01)	0.06
BMI (%)			
Normal	27.27	48.65	
Overweight	48.48	24.32	0.09
Obese	24.24	27.03	
Caloric intake (kcal) (mean, SEM)	2053.78 (149.9)	2104.89 (252.46)	0.86

TABLE 1.2: 454 dataset characteristics before and after QC for RDP and OTU pipelines

RDP Pipeline	Original	After QC
Total # of Sequences	600354	598645
Average/Sample	8455.69	8431.62
SD	3840.73	3843.29
Average Sequence Length	343.131	343.575
OTU Pipeline	Original	After QC
Total # of Sequences	600354	532506
Average/Sample	8455.69	7500.08
SD	3840.73	3578.55
Average Sequence Length	343.131	302.034

CHAPTER 2: FILTERING OUT BATCH-EFFECTS IN METAGENOMIC ANALYSIS REVEALS A TRUE BIOLOGICAL SIGNAL

2.1 Abstract

Difference between populations, different body sites and disease states has been the focus of many metagenomic studies, including the Human Microbiome project. As with other comparative studies, caution needs to be exerted in these studies to separate real biological differences from technical artifacts. Using an example of a major batch- effect that we discovered during the analysis phase of our colorectal cancer project (chapter 1), we illustrate how filtering out batch-effects helped us to reveal a very important biological result in our data.

2.2 Background and significance

The quality and validity of results obtained from any biological research, including research involving high-throughput technologies like microarrays, mass spectrometry and sequencing requires quality control measures to be used during the design, experimental and analysis phases of the research process. During the course of a metagenomic study such as the one described in chapter 1, a series of experimental methods, protocols, hardware, software and analyses are used. Keeping all these conditions constant is essentially impossible. Batch-effects occur when the outcome of experiments is affected by the group in which the samples are processed. Batches can be either reagent batches,

date batches, or technician associated batches. For example, batch- effects may occur if a subgroup of samples were processed in one lab versus the other or by one technician versus the other. Batch-effects are important technical artifacts commonly encountered in many metagenomic and genomic studies; they must be accounted for in order to reap the benefits from these studies. Low throughput techniques such as Western blotting and PCR are also prone to batch-effects but batch-effects are much more easily detected in high-throughput methods like microarrays, sequencing (454, Illumina etc.) and proteomics [91]. Also due to the fact that high-throughput experiments are generally performed in larger scale they are processed in different locations, on different dates, and possibly by various technicians in order to distribute workload. All of these factors make high-throughput studies, like metagenomic studies extremely vulnerable to batch-effects.

Studies that demonstrated the correlation between biological variables and technical variables have been reported in literature[111],[112] and these studies acknowledge the fact that batch-effects are critical in high throughput analyses and have to be dealt with in order to reach biologically accurate conclusions. In this chapter we illustrate, through our own dataset, how batch-effects masked true biological effects and how by filtering them out we were able to salvage the study.

2.3 Materials and Methods

2.3.1 Methods

Study Participants, colonoscopy and Biopsy procedures and DNA Extraction were as described in the previous chapter.

2.3.2 Bacterial Identification

The first step in the data analysis process involved a preliminary QC filter (downstream of the filters from the Roche-454 GS-FLX software). We removed sequences from our dataset if (1) there were any Ns in the sequence or the 5' primer did not exactly match the expected 5' primer or if the average quality score was less than 20. We then removed the 5' primer sequence from our reads that have survived above filtering. Only trimmed filtered sequences with a length between 200-500bp were kept in our data set and submitted to the RDP classifier algorithm 2.0[46] for taxonomic identification at various taxonomic levels. Sequences assigned in each sample to various taxa, from phylum level up to genus level, were counted at the RDP confidence threshold of 80%. Raw counts were normalized, and then log transformed using the normalization scheme mentioned below, before proceeding further.

LOG10 ((Raw count / # of sequences in that sample) +0.001)

Only taxa with >= 10seqs in at least 25% of the samples were selected for downstream statistical analyses.

2.3.3 Statistical analyses

The Shannon-Wiener Diversity Index, H, was calculated using the following equation: $H = -\sum P_i (\ln P_i)$ where P_i is the proportion of each species (taxa) in the sample. Student's t-tests were performed to compare the mean similarities of the groups, case and control. Student's t-tests, Wilcoxons, PCA and hierarchical clustering were performed using JMP 8.0 (SAS Institute, Cary NC).

2.4 Results and Discussion

2.4.1 Descriptive characteristics of study participants

We analyzed the adherent microbiota from mucosal biopsies from 167 individuals, including 80 adenoma cases and 87 non-adenoma controls based on the16S rDNA genes and 454 titanium pyrosequencing methods. Case subjects were slightly older (case-57.2 years) compared to controls (55.5 years). Cases were more likely to have higher Waist-to-Hip-Ratio than controls (p=0.0001) and be overweight or obese (p=0.018). There were no significant differences between cases and controls for smoking, fiber intake, caloric intake, and fat (Table 2.1). After applying a quality filter (see methods), a total of 1,411,767 sequences were present and of these, 1,407,099 were classified as domain Bacteria at a confidence threshold of 80% by the RDP classification algorithm[46]. The average number of sequences/subject was ~8400 (8403.37±3133.38) and the average sequence length was ~350bp (341.37±86.9).

2.4.2 All samples clustered into two distinct groups

We started our analysis of this dataset with an unsupervised approach by asking whether all samples from the study form natural groups with respect to their microbiome composition, independent of metadata associated with each sample. Principal component analysis of the log normalized abundance of all taxa at the genus level revealed 2 distinct clusters (Figure 2.1). The samples in cluster 1 showed a very different microbial profile compared to the samples in cluster 2 (Tables 2.2 and 2.3). The cluster 1 had a lot of within-cluster variability with significant differences in microbial abundance between cases and controls (Table 2.2), whereas cluster 2 was compact with very little variability between all the samples within the cluster. Most of the case subjects belonged to this cluster and there were no significant differences between the case and control subjects within this cluster (Table 2.3). Due to the fact that from the PCA all samples in cluster 2

were incredibly similar to one another, with respect to their microbiome composition, and that previous studies [113],[59],[52] had suggested that each individual had a unique microbiome fingerprint, we suspected that there could be some technical artifact that was causing the incongruent pattern in the samples belonging to cluster 2.

2.4.3 The distinct clustering was due to a batch-effect

To check if our notion was indeed true, we looked for a correlation between these naturally occurring groups and the metadata associated with the samples to see if any of the metadata categories were responsible for this separation. Just as we had expected, our results indicated that there is an almost perfect correlation of clusters 1 and 2 with the technical variable groups (batches) namely the “Date sent for Sequencing” and “Location of Stock DNA” (Figure 2.3). Once we had confirmed that this behavior was due to a batch- effect, the next step was to find a way to get rid of the batch-effect, if possible. The question was, given the two distinct batches, is there a reason to believe that one of them is biologically “correct”? If so, which batch is the biologically correct batch and which one is the incorrect batch?

2.4.4 Batch-1 had a biological signature

Based on previous literature [69], [7] we would expect a true gut microbiome sample to have certain broad characteristics. Firstly, as mentioned above, pioneering studies in the field of human microbiome res**e**arch [113], [52] have suggested that each person is like a “microbial island”, with respect to their gut microbial composition, meaning that each would have their own microbiome signature. From our PCA (Figure 2.2) and from comparison of the Shannon Diversity index of batches 1 and 2 (Figure 2.3), it is obvious that samples in batch-1 conformed to this pre-existing knowledge but the samples in

batch-2 were too similar to one another to satisfy the “individual microbiome concept”. In addition, previous literature in this field suggests that the composition of the gut microbiota in human and most mammals is dominated by the two phyla [114], [115], Bacteroidetes and Firmicutes and the overall percentage of Bacteroidetes + Firmicutes (B+F) in the samples is expected to be about 90% of the total gut bacteria. From figure 2.4 it is evident that samples in batch-1 meet that expectation whereas the samples in batch-2 had B+F percentages that are considerably lower than we would expect based on previously published human gut microbial datasets[70], [115] including a dataset from our own lab[116]. Based on this justification, the 95 samples that belong to batch-2 were removed from further analysis and only 71 samples were further analyzed to look for a microbial signature associated with colorectal adenoma status.

The results of our analysis, after removal of the batch-effect, provide another level of justification for our decision to exclude the samples in batch-2 from further analysis (Figure 2.5, Chapter 1 Appendix A, Supplementary figures 1 and 2 and Appendix A, Supplementary Tables 1 and 2). When all of our samples (including the batch-effect samples) were included in our analysis, we found that no taxa at Phylum level and only 2 taxa at genus level were significantly different between the cases and controls at 10% FDR (Tables 2.4 and 2.5). But once we filtered out the batch-effect, our results improved and we now have 3 phyla and 31 genera that are significantly different between case and control at 10% FDR threshold (Appendix A, Supplementary Tables 1 and 2). This clearly indicates that the biologically correct signature (Figures 2.3 and 2.4) of the samples in batch-1(that helped us make the decision that it is the good batch) is also linked to significant differences between the case vs. control samples in that batch. By

using prior knowledge in the field and with the help of bioinformatic analysis tools, we were able to “save” our study and recover the true biological potential of our data from getting lost in the technical noise.

To summarize, while we were able tell the “biological” signal apart from what is likely non-biological noise, the exact reason for the deviation of the samples belonging to the affected batch cannot be pin-pointed, since the two technical variables stock DNA (hallway freezer vs. lab freezer) and the date sent for sequencing (09/30/09 vs. other) are 100% confounded with each other. Fortunately, since these technical variables were not confounded with the biological variable (disease status, case and control) of interest to us, we were able to successfully detect the batch-effect and remove it from our dataset to reveal the important biological effect in our study. This chapter thus demonstrates both the necessity and feasibility of examining batch-effects in metagenomic datasets and provides a possible analysis path for detecting such artifacts and removing them.

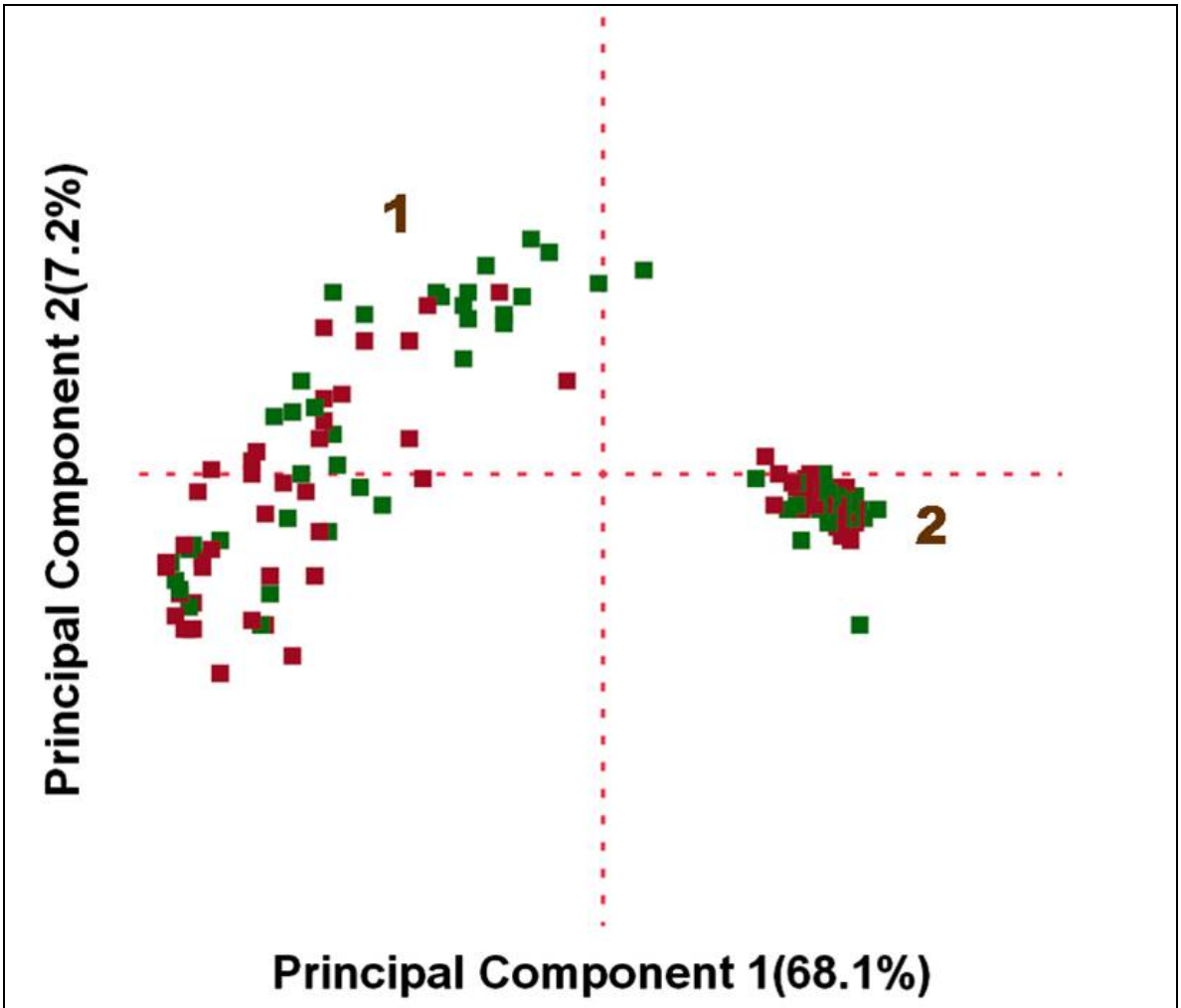


FIGURE 2.1: Principal component Analysis (PCA) on the normalized log abundances of all taxa at the genus level shows 2 distinct clusters that do not correlate with disease status; Case (red), Control (green).

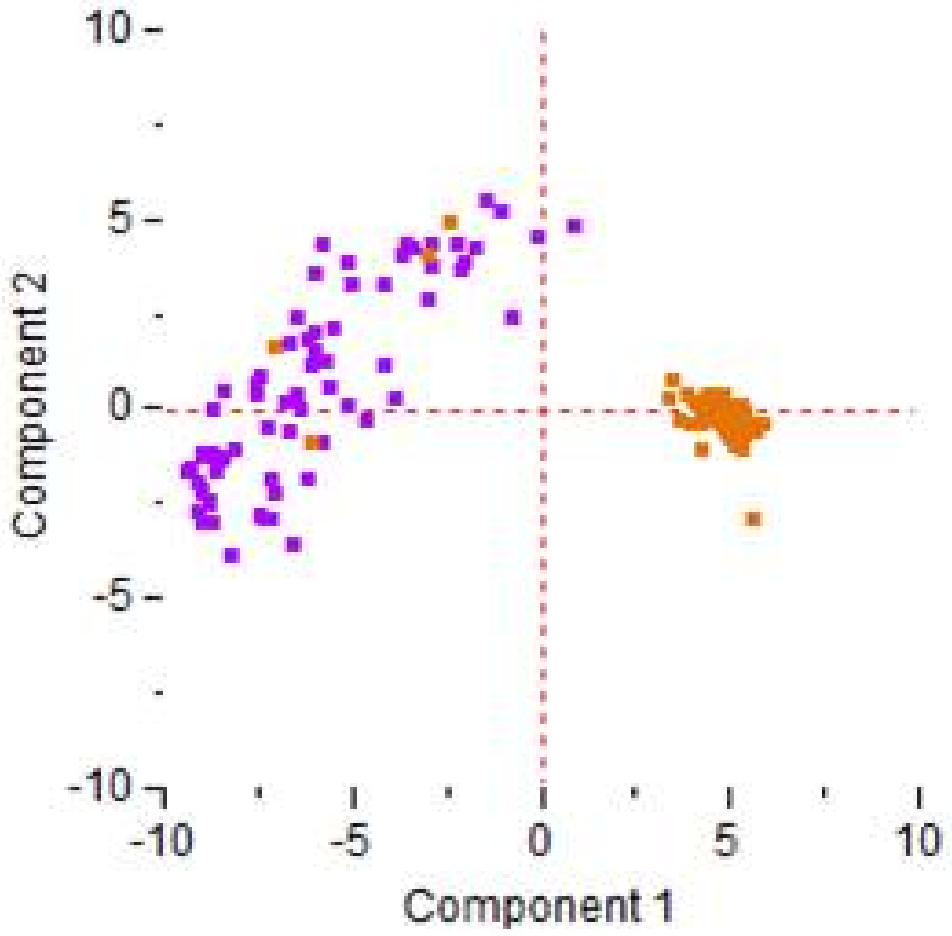


FIGURE 2.2: Principal component Analysis (PCA) on the normalized log abundances of all taxa at the genus level shows 2 distinct clusters that correlate almost perfectly with technical variables; Date sent for Sequencing (Purple: Other; Orange: 09/30/09) and Location of Stock DNA (Purple: Hallway freezer; Orange: Lab freezer) indicating a batch-effect.

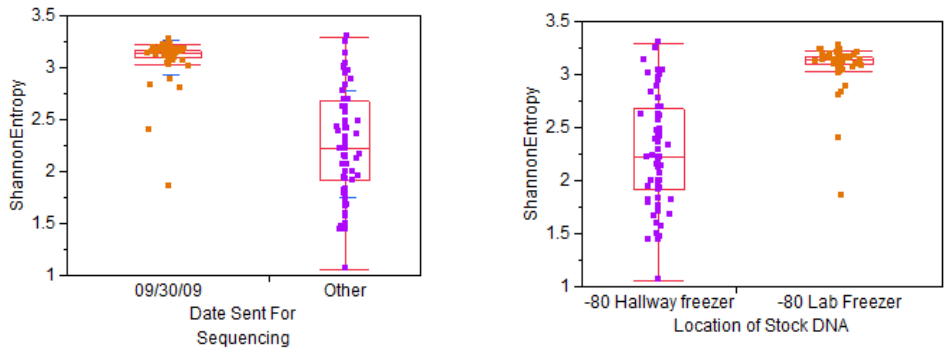


FIGURE 2.3: Wilcoxon test ($p<0.001$) on the Shannon Diversity indices between the technical variable groups (Date sent for Sequencing and Location of Stock DNA) shows that most samples in batch-2; Date sent for sequencing (09/30/09) and Location of Stock DNA (-80 Lab freezer) have unusually similar Shannon diversity indices.

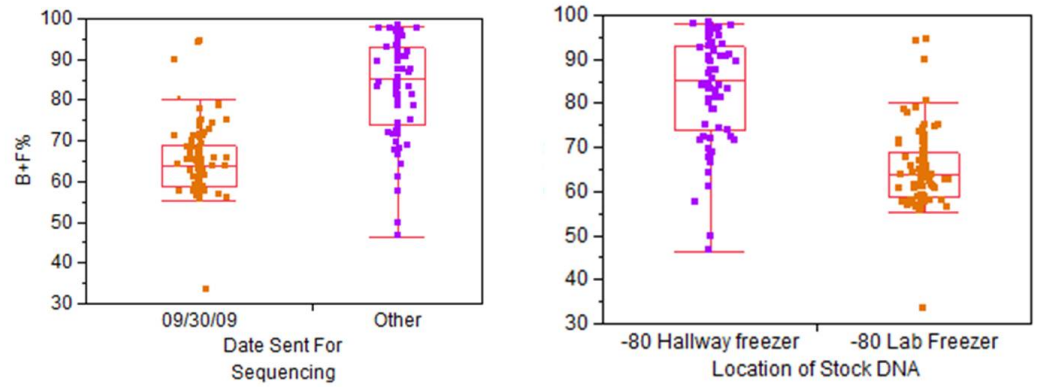


FIGURE 2.4: Wilcoxon test ($p<0.001$) on the percentage of Bacteroidetes + Firmicutes between the technical variable groups (Date sent for Sequencing and Location of Stock DNA) show that the most samples in batch-2; Date sent for sequencing (09/30/09) and Location of Stock DNA (-80 Lab freezer) have an abnormally low B+F%.

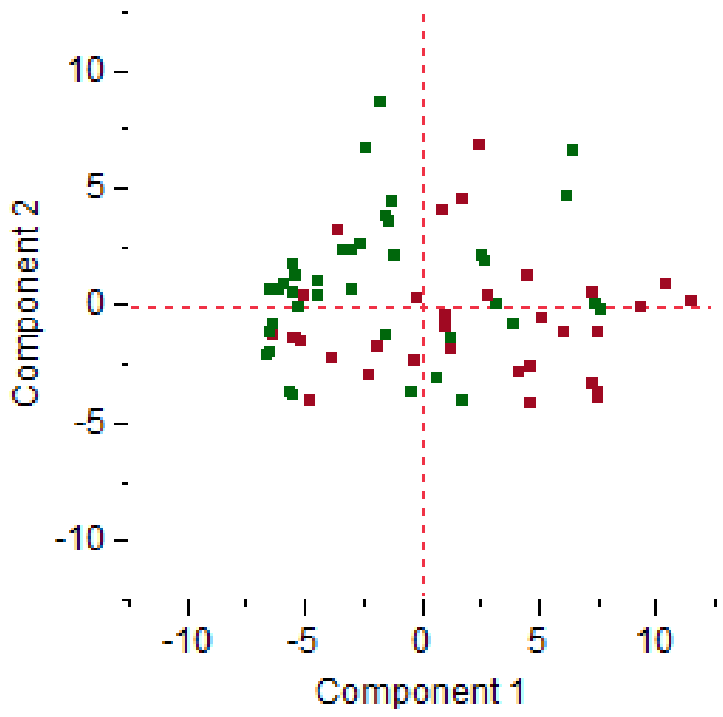


FIGURE 2.5: Principal component Analysis (PCA) on the normalized log abundances of all taxa at the genus level after removal of the batch-effect samples; Case (red), Control (green).

TABLE 2.1: General characteristics of the study participants, cases (80) and controls (87). p-Values are based on t-tests between case and control (age, WHR and caloric intake) or the Chi square test (% Male and %BMI).

Characteristic	Case n=80	Control n=87	p-values
Age (mean, SD)	57.2 ± 6.88	55.5 ± 6.09	0.099
Male (%)	58%	38%	0.01
Family History of CRC (Yes %)	4%	2%	NA
Waist-hip ratio (mean, SD)	0.94 ± 0.076	0.89 ± 0.079	0.0001
BMI (mean, SD)	27.34 ± 4.53	26.6, 5.94	0.369
Smoking (Yes %)	56%	49%	NA
Calories (mean, SD)	2041.51 ± 800.71	1976.13 ± 1062.03	0.66
Alcohol_g (mean, SD)	11.46 ± 16.21	16.99 ± 62.97	0.46
Total_fat_g (mean, SD)	75.88 ± 31.22	73.72 ± 31.75	0.67
Dietary Fiber (mean, SD)	20.09 ± 8.93	20.71 ± 10.21	0.68

TABLE 2.2: T-tests on log-normalized abundances of genera in cases vs. controls in cluster 1. Only top 10 taxa based on p-Values shown. Significant differences between case and control seen at 10% FDR. T-test p-Values were corrected for multiple testing using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted Rank of the taxon.

Taxa	p-Value	Rank	n*p/R
Helicobacter	0.000118336	1	0.005680125
Acidovorax	0.000209127	2	0.005019045
Lactobacillus	0.000500604	3	0.008009666
Cloacibacterium	0.000510639	4	0.006127667
Lactococcus	0.000550592	5	0.005285679
Stenotrophomonas	0.000921315	6	0.007370519
Turicibacter	0.001653894	7	0.011340985
Weissella	0.001660807	8	0.009964842
Delftia	0.001994309	9	0.010636313
Acinetobacter	0.002363268	10	0.011343687

TABLE 2.3: T-tests on log-normalized abundances of genera in cases vs. controls in cluster 2. Only top 10 taxa based on p-Values shown. No significant differences between case and control seen at 10% FDR. T-test p-Values were corrected for multiple testing using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted Rank of the taxon.

Taxa	p-Value	Rank	n*p/R
Pantoea	0.03707127	1	1.77942096
Burkholderia	0.045751861	2	1.098044653
Dorea	0.053883447	3	0.862135153
Turicibacter	0.061158934	4	0.733907206
Bacillaceae_1	0.069582119	5	0.667988338
Lactococcus	0.081598587	6	0.652788692
Parabacteroides	0.126241514	7	0.865656094
Chryseobacterium	0.143720051	8	0.862320304
Bryantella	0.153965999	9	0.821151995
Streptococcus	0.165889593	10	0.796270046

TABLE 2.4: T-tests on log-normalized abundances of phyla in cases (80 subjects) vs. controls (87 subjects), before removing batch-effect shown. Only phyla which have at least 10 sequences assigned to them in 25% of the samples are shown. T-test p-Values were corrected for multiple testing[95] using (n*p)/R where n= total number of taxa tested, p= raw p-Value and R= sorted Rank of the taxon.

Taxa	t-Test p_Value	RANK	n*p/R
Firmicutes	0.019160329	1	0.114961971
Cyanobacteria	0.040488222	2	0.121464665
Actinobacteria	0.073114722	3	0.146229443
Proteobacteria	0.155797173	4	0.233695759
TM7	0.388303816	5	0.46596458
Bacteroidetes	0.532809351	6	0.532809351

TABLE 2.5: T-tests on log-normalized abundances of genera in cases (80 subjects) vs. controls (87 subjects), before removing batch-effect shown. Only genera which have at least 10 sequences assigned to them in 25% of the samples are shown. T-test p-Values

were corrected for multiple testing[95] using (n*p)/R where n= total number of taxa tested, p= raw p-Value and R= sorted Rank of the taxon.

Taxa	t-Test p_Value	RANK	n*p/R
Cloacibacterium	0.000577837	1	0.027736174
Acidovorax	0.002426601	2	0.058238416
Acinetobacter	0.011368024	3	0.181888381
Streptococcus	0.016895792	4	0.202749502
Lactobacillus	0.068901773	5	0.661457017
Bacillaceae_1	0.084901022	6	0.679208178
Helicobacter	0.10190281	7	0.698762128
Sutterella	0.105286728	8	0.63172037
Delftia	0.13381527	9	0.713681438
Micrococcineae	0.134231434	10	0.644310882
Stenotrophomonas	0.150206126	11	0.655444915
Dorea	0.153808383	12	0.615233532
Sphingobium	0.156600567	13	0.578217477
Pantoea	0.169016522	14	0.579485218
Sphingomonas	0.174709512	15	0.559070438
Alistipes	0.195506126	16	0.586518378
Exiguobacterium	0.204798818	17	0.578255485
Lactococcus	0.212610872	18	0.566962325
Bryantella	0.212887223	19	0.537820352
Chryseobacterium	0.220024908	20	0.528059779
Turicibacter	0.259646746	21	0.593478277
Pseudomonas	0.280269635	22	0.611497385
Agrobacterium	0.36124946	23	0.753911917
Serratia	0.370766763	24	0.741533525
Rikenella	0.399816152	25	0.767647011
Leuconostoc	0.402214518	26	0.742549879
Weissella	0.446325861	27	0.793468197
Coprococcus	0.455961712	28	0.78164865
Burkholderia	0.456934521	29	0.756305414
Roseburia	0.514954353	30	0.823926965
Shinella	0.533258234	31	0.825690169
Ruminococcus	0.593612066	32	0.890418099
Subdoligranulum	0.597673669	33	0.869343518
Methylobacterium	0.646389809	34	0.912550319
Anaerotruncus	0.6491764	35	0.890299063
Flavimonas	0.655266685	36	0.873688913

Bacteroides	0.692484055	37	0.898357693
Variovorax	0.704272198	38	0.889606986
Chryseomonas	0.721439112	39	0.887925061
Peptostreptococcaceae_Incertae_Sedis	0.785155542	40	0.942186651
Faecalibacterium	0.792620103	41	0.927945486
Erwinia	0.839482176	42	0.959408201
Coriobacterineae	0.871637423	43	0.972990612
Clostridiaceae_1	0.873337628	44	0.952731958
Coprobacillus	0.913262206	45	0.974146353
Erysipelotrichaceae_Incertae_Sedis	0.952178019	46	0.993577064
Parabacteroides	0.962042355	47	0.982511342
Lachnospiraceae_Incertae_Sedis	0.965972327	48	0.965972327

CHAPTER 3: MOLECULAR DIVERSITY OF A NORTH CAROLINA WASTEWATER TREATMENT PLANT AS REVEALED BY PYROSEQUENCING [117]

3.1 Abstract

We report the results of pyrosequencing DNA collected from the activated sludge basin of a wastewater treatment plant in Charlotte, North Carolina, U.S.A. Using the 454-FLX technology, we generated 378,601 sequences with an average read length of 250.4 base pairs. Running the 454 assembly algorithm over our sequences yielded very poor assembly with only 0.3% of our sequences participating in assembly of significant contigs. Of the 117 contigs greater than 500 base pairs that were assembled, the most common annotations were to transposases and hypothetical proteins. Comparing our sequences to known microbial genomes showed non-specific recruitment indicating that previously described taxa are only distantly related to the most abundant microbes in this treatment plant. A comparison of proteins generated by translating our sequence set to translations of other sequenced microbiomes shows a distinct metabolic profile for activated sludge with high counts for genes involved in metabolism of aromatic compounds and low counts for genes involved in photosynthesis. Taken together, these data document the substantial levels of microbial diversity within activated sludge and further establish the great utility of pyrosequencing for investigating diversity in complex ecosystems.

3.2 Background and significance

The entire biosphere is influenced by the ability of microorganisms to transform the world around them. Microbes have the ability to convert the some of the important elements of life like carbon, nitrogen, oxygen, and sulfur, from their inaccessible natural forms to simpler forms to make them available to other living beings. Microbes in their role as scavengers help clean up the both organic (biodegradable wastes) and inorganic (chemical and oil spills) wastes from the environment [118], [119], [120]. While some of these activities are carried out by individual microbes, most of these processes are mediated by complex microbial communities that have the ability to adapt quickly to the changes in their surrounding environment. One of the environments where microorganisms play a critical role is within wastewater treatment plants [121], [122], [123]. Wastewater treatment plants are probably the largest "microbially-mediated biotechnology processes" on the planet[124] and they play an very important role in maintenance of public health.

Although largely invisible in the urban landscape when they are functioning well, wastewater treatment plants are integral to the municipal obligation to protect public health, aquatic ecosystems, and the quality of life. At the heart of wastewater treatment plants is a process whereby a dense microbial consortium is employed to remove organic and nutrient contaminants. These microbes used to treat wastewater are a crucial tool in environmental protection. The current use of molecular techniques that do not require the isolation and cultivation of microorganisms[125-126], including 16S rRNA[127-129] and fluorescent in situ hybridization[130] have greatly expanded our understanding of wastewater microbial communities. Researchers have identified many bacteria of

importance to wastewater treatment including those involved in biological phosphorus removal [131-133] nitrifiers[130, 134-135], denitrifiers[123, 136-137] and methanogens [138-139]. Molecular techniques have also improved our understanding of fundamental processes such as nitrification and denitrification as well as plant upsets, such as foaming [140-141], which can decrease treatment efficiency.

In this Chapter, we apply pyrosequencing technology to probe the molecular diversity of the aerobic basin of a wastewater treatment plant in Charlotte, North Carolina, U.S.A. In line with other studies of complex microbial communities [10, 142], we observed astounding levels of diversity. We find that the most prevalent microbes in the wastewater treatment plant have substantial regions of their genomes that are poorly described by existing sequence databases. Our results demonstrate that despite recent technological advances that allow for the identification of microorganisms, the microbial population of wastewater treatment plants remains under sampled and inadequately characterized. During the course of this study we also introduce the various bioinformatic methods used in the metagenomic analysis of complex ecosystems and discuss the advantages as well as the limitations of some of these methods. Our results are a first step towards a more complete molecular characterization of this important but understudied microbial community.

3.3 Materials and Methods

The Mallard Creek Water Reclamation Facility is located in Charlotte, North Carolina. The plant has an average daily inflow of 7.5 million gallons and the wastewater is mostly domestic, with additional input from the University of North Carolina Charlotte, University City Carolinas Medical Center hospital, and several

industrial users. A schematic of the flow through the plant is shown in Supplemental Figure 1. Influent raw wastewater is screened and sent through grit removal before it is routed to day tank equalization basins that distribute the flow among three primary clarifiers. Primary effluent enters anoxic basins, where it is joined by recycle flow from the aeration basins. Effluent from the anoxic basins enters aeration basins (solids retention time ~ 8 days) and then flows to secondary clarifiers. Clarified effluent is routed to denitrification filters and then to UV disinfection before discharge to Mallard Creek.

The plant NPDES (National Pollutant Discharge Elimination System) permit requires the plant to meet a monthly CBOD5 of 4.2 mg/L in the summer and 8.3 mg/L in the winter months. Ammonia nitrogen (NH3-N) levels must be below 1 mg/L and 2 mg/L in summer and winter, respectively. There are no other nitrogen or phosphorus limits. Total suspended solids are limited to a maximum of 30 mg/L, and the pH must be between 6 and 9 standard units. Fecal coliforms counts must be less than 200 colony forming units (cfu) per 100 mL sample. These limits are routinely met by the plant unless there are extreme weather events or plant upsets. Wastewater entering the secondary treatment system was monitored over a six month period for filtered flocculated COD, a good estimator of readily biodegradable soluble organics, and values ranged from 40-75 mg/L. Ammonia nitrogen concentrations in this same flow ranged from 12-24 mg/L, with the concentration varying in part due to return flow from digested sludge dewatering.

On the morning of March 20, 2007 we collected a 50 mL sample from the aeration basin using a plastic dipper. At the time of sample collection, temperature in the aeration basin was 18.5^{o}C and pH was 6.5. The sample was decanted to remove as much foam as

possible before transferring the liquid to a sterile tube. DNA was extracted from the sample using a Mo Bio UltraClean Water DNA Kit. The sample tube was inverted several times to maximize homogeneity and a 10 mL aliquot was removed and pipetted on to the provided filter (0.22 µm). Filtrate was discarded and DNA was extracted from the membrane using the manufacturer's protocol. The final DNA extract was analyzed for purity and concentration using a NanoDrop ND-1000 spectrophotometer. Approximately 100 µl of extracted DNA was concentrated in a speed vac and resuspended in about 12 ul of molecular grade biology water. The final sample concentration was 479 ng/µl as determined by a NanoDrop spectrophotometer. Preliminary analysis of the DNA using Denaturing Gradient Gel Electrophoresis (DGGE) indicated substantial diversity in the observed bands confirming that our DNA extraction was successful (data not shown). The sample was submitted to 454 Life Sciences for pyrosequencing of the 454-FLX platform. The methodology underlying pyrosequencing has been documented elsewhere [25].

Sequences and quality scores from our pyrosequencing run have been submitted to the NCBI short read archive (accession numbers SRA001012). All the supplemental material related to this chapter can be found at; http://aem.asm.org/cgi/content/full/75/6/1688/DC1?maxtoshow=&hits=10&RESULTFORMAT=&fulltext=Nina+Sanapareddy&searchid=1&FIRSTINDEX=0&resourcetype=HWCIT.

3.4 Results and Discussion

3.4.1 Our sequence set largely fails to assemble, although contigs that were generated from the assembly include many transposons and hypothetical proteins.

Our pyrosequencing run yielded 378,601 sequences with an average read length of 250.4 ± 29.1 (mean ± SD). The distribution of sequence lengths was approximately normal with a small left tail indicating some short reads (Supplemental Figure 3, see methods). We attempted to assemble sequences in this dataset using version 1.1.02 of the GS De Novo Assembler of the Genome Sequencer FLX Data Analysis suite with the default parameters applied. This assembly algorithm attempts to combine individual sequence reads into longer "contigs". Given that metagenomic datasets of complex ecosystems have been extremely resistant to assembly [142-143], we expected to see very little assembly in our dataset. The 454 sequence assembler defines a "large" contig as one that consists of at least 500 base pairs. Because our average sequence length was ~250 base pairs, this threshold could be achieved with the overlap of a modest number of our sequences. Despite this, only 1154 (or approximately 0.3%) of our reads were recruited into 117 contigs greater than 500 base pairs (the sequences of these contigs are available as Supplemental File 1, see methods). To assign possible functions to these contigs, we used the GenMark algorithm[144] to predict genes on our contigs and then performed a BLASTP search of these predicted proteins against the Pfam database. This method produces more assignments than other approaches including those based on profile searches (Supplemental File 11, see methods). With an e-score cutoff of 0.01, this approach found matches for 75% (88/117) of our large contigs (Supplemental File 2, see methods). Of these matches, 22% (20/88) were to hypothetical proteins and 21% (19/88) were to transposases. The prevalence of transposases in our assembled contigs strongly suggests that transposons are much more strongly conserved across metagenomes than other genomic regions while the prevalence of hypothetical proteins shows that the

function of many of the highly conserved regions of our metagenome is poorly understood.

This failure of the 454 assembly algorithm to assemble 99.7% of our sequence reads emphasizes the great diversity of the microbial community within the treatment plant. Because previous literature has found a similar failure of assembly algorithms on metagenomic communities characterized by Sanger sequencing[142], as well as on simulated data sets created by Sanger sequencing reads[143], we would not expect a significantly improved degree of assembly even if our sequence reads were longer.

3.4.2 The majority of taxa in the wastewater treatment plant cannot be classified at the Genus level.

In order to discover the 16S rRNA genes within our dataset, we downloaded the 16S rRNA gene FASTA DNA sequences from v. 9.52 of the Ribosomal Database Project (RDP)[145] and used these sequences to create a BLAST database. Using the blastn algorithm, we asked which of our 378,601 query sequences could be found in this RDP database with an e-score of e<=0.01 (Supplemental File 11, see methods). The resulting 648 sequences (available as Supplemental File 3, see methods) were run through the RDP classification algorithm[46] (see supplementary methods). The RDP classifier algorithm uses Bayesian statistics to assign taxa to 16S rRNA gene sequences. The output of this algorithm includes a confidence score, which ranges from 0 to 100, that indicates the degree of confidence that can be assigned to the classification based on the results of 100 bootstrap trials (see[46] for more details). The recommended threshold for assigning of a taxa by the RDP algorithm is a confidence score >=80. Because sequence reads as short as 90 basepairs have been shown to suffice to accurately characterize taxa [32, 146], we

anticipate that our results would not be substantially different even if we had a read length longer than 250 basepairs.

The classifications of the 148 16S rRNA sequences that could be assigned to Phylum with a confidence score of >=80 are provided as Supplemental File 4(see methods) and are summarized in Figure 3.1. In another paper[147], we show that these classifications of 16S rRNA sequences derived from the whole-genome wastewater sequence set are well correlated with results from PCR experiments targeting the 16S rRNA gene. At the Phylum level, the observed taxa are dominated by the Proteobacteria with ~70% of the classifiable taxa belonging to this category (Figure 3.1 top panel). Moving from Phylum to Genus, fewer of the sequences can be classified with an RDP confidence score of at least 80%. At the Genus level, nearly 60% of the sequences cannot be classified at a RDP threshold of 80 and, of the taxa that can be classified; there is no dominant taxon (Figure 3.1). These data demonstrate the extraordinary microbial diversity of activated sludge and is consistent with reports from other complex environments [43, 142, 148]. We note that the inability of the RDP algorithm to classify these sequences to taxa with high confidence is not primarily the result of our 16S rRNA sequences having never been previously observed. Figure 3.2 shows that many of the sequences with RDP scores <80% (to the left of the vertical lines) have very high percent identities to previously described sequences. These results demonstrate that for wastewater treatment plants, as is the case for other complex ecosystems, the accumulation of 16S rRNA sequences in public databases is vastly outpacing our ability to classify them, and that this problem becomes more pronounced as one moves from Phylum towards Genus. Presumably future annotation efforts will rectify this problem.

3.4.3 16S rRNA gene sequences from freshwater, soil and other wastewater studies dominate our sequence set.

For each of the 648 sequences in our pyrosequencing dataset that matched the 16S RDP database (v 9.52) at an e-score cutoff of <=0.01, we manually annotated where the corresponding RDP sequence was discovered. This was done by manual inspection of the Genbank records for these 648 sequences. The results of this annotation can be found in Supplemental File 8 (see methods) and are graphed in Figure 3.3. The x-axis of Figure 3.3 indicates our classification while the y-axis indicates the e-score with which the top hit from each of our query sequences matched the RDP database. We see that while a large number of environments had at least one hit, if we restrict ourselves to environments with multiple hits at high stringency (i.e. low e-score), only three environments are well represented: freshwater, soil and other wastewater studies (Figure 3.3). While, of course, the low number of sequences for some of the other environments may simply reflect the low number of sequences from that environment in the 16S RDP database, there is a strikingly small number of sequences with high scores that relate to two 16S populations that are well represented in the database: marine and human. The relatively small number of human-derived 16S rRNA sequences observed is particularly interesting given the vast number of human microbes deposited into the wastewater treatment plant each day. These results show that the environment within the wastewater treatment plant exhibits strong selection pressure against the microbes that are present in human feces.

3.4.4 Sequenced bacterial genomes are not well represented in the wastewater metagenome.

When using BLASTN to compare our sequences to the nt database, only 34% (73,274/378,601) match the nt database even at a relaxed threshold e-score cutoff of e<0.01 (data not shown; see Supplemental Bioinformatics methods). Of the sequences that do match the nt database at this threshold, the vast majority (over 98%) have their best hit to bacteria taxa (data not shown; see Supplemental Bioinformatics methods for details). Since wastewater treatment plants are known to harbor many eukaryotes (e.g. [149]), this result likely reflects our DNA isolation strategy, which was designed to capture prokaryotic DNA, rather than the "true" ratio of prokaryotes to eukaryotes in the treatment plant.

As of November 2008, there are 772 complete bacterial genomes at the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz). In order to explore how well these known genomes are represented in the treatment plant, we used BLASTN to compare our wastewater sequences to the 1,442 assembled genome and plasmid sequences from the 772 sequenced bacteria. In order to eliminate spurious hits, we required that any hit matched at least 75 nucleotides in our query sequence (see Supplementary Bioinformatics methods for more details). Because our average sequence length was 250.4 basepairs (Supplemental Figure 3, Supplemental File 3, see methods), this is not an overly conservative criterion. Under this criterion, only 20% (73,274/378,601) of our sequences matched to any of the known bacterial genomes. This result again reflects the great diversity of the wastewater treatment plant and emphasizes a key challenge for genomics; despite the considerable effort that has been expended in the microbial genome projects, the great majority of our sequence reads are not found in known genomes.

For the sequences that do match to known genomes, we can determine how closely the sequenced genomes from cultivated organisms match the genomes present in our wastewater metagenome. We calculated for each of the 1,442 assembled sequences from the 772 finished genome projects the number of nucleotides in that genome that have a BLASTN match that is aligned to at least one of our wastewater sequences. Dividing this number by the total length of each assembled sequence yields the "fraction genome covered". Figure 3.4. shows that even the most well represented assembled genome, the nitroaromatic compound-degrader Acidovorax sp. JS42 (NC_008782), has a match to only 25% of its sequence to our wastewater metagenome. Table 3.1 shows that the fraction genome covered is similarly poor for the ten genomes that recruited the most reads from our wastewater metagenome.

Figure 3.5 shows a recruitment graph of our wastewater treatment plant for the assembled Acidovorax sp. JS42 chromosome, which recruited the most sequences from our wastewater genome (Table 3.1; Supplemental File 6, see methods). On the x-axis is the position where sequence reads are mapped with blastn against the Acidovorax genome. On the y-axis is the percent identity of the read when compared to the matching subsection of the Acidovorax genome. Figure 3.5 shows sequences from two different sources: our March 20 aeration basin pyrosequencing run (black lines) and the environmental sequence database from NCBI downloaded in June 2007 (red lines), which at that time was largely dominated by sequences from the J. Craig Venter Institute's Global Ocean Sampling (GOS) [142]. We included the environmental sequence database because we wanted to assess how specific our wastewater treatment plant results were relative to other metagenomic sequencing databases.

The pattern seen in Figure 3.5 is typical of non-specific recruitment. For regions of the genome with conserved genes, both sources of sequence matched to the genome, but the percent identities were usually below 90%. For regions of the genome that are poorly conserved, such as the putative transmembrane protein (marked by arrows in the annotation section at the top of Figure 3.5), very few sequences from either source mapped to the genome. We observed similar patterns of non-specific recruitment over a number of the genomes that recruited large numbers of sequence reads in our March 20 aeration basin dataset (data not shown). In the Venter GOS survey, a similar pattern of non-specific recruitment was observed against nearly every known microbial genome despite the presence of over 7,600,000 sequences in the dataset [142]. This result is one of the principle reasons that the GOS study concluded that microbial diversity in the oceans is profound [142]. Our results show that the most abundant microbes in the wastewater treatment plant have genomes that are largely uncharacterized. Moreover, the pattern of non-specific recruitment shown in Figure 3.5 suggests that even additional whole-genome shotgun sequencing would not improve the match between known genomes and the sequences observed in our metagenome.

One genome of particular interest that is not yet deposited as an assembled genome at NCBI is the “Candidatus Accumulibacter phosphatis” taxa that dominates two lab scale EBPR sludges recently sequenced[124]. Although the assembled genome of this taxa has not yet been publicly released, we saw similar patterns of non-specific recruitment to the largest assembled contigs that have been released as we saw to the publically available assembled genomes (data not shown). This suggests that the “Candidatus

Accumulibacter phosphatis" taxon is not a dominant member of our North Carolina wastewater genome.

The great diversity of our wastewater metagenome caused very few contigs to be recruited. Of the sequences that were recruited to contigs, a substantial fraction involved transposases. We might expect, therefore, a different pattern of recruitment around transposons. Figure 3.6 shows a region of the Acidovorax genome around a transposase with a stark exception to the pattern of non-specific recruitment. A large number of sequences from our metagenome recruited to this region with a nearly perfect match. Interestingly, a number of marine sequences from the Global Ocean Survey[142] also matched the region around this transposase (red lines), suggesting that, unlike most genomic regions, parts of this transposon are conserved across a wide environmental space.

3.4.5 When mapped to protein space, the wastewater metagenome displays a distinct metabolic profile.

By translating our nucleotide sequences in all six frames and mapping the translated sequences to known proteins, we can generate a distinct metabolic profile for our wastewater sequences. This approach, asking which genes a microbial community is capable of producing, has been successfully used to analyze the metabolic signatures of a number of metagenomic sequence sets [150-151]. To perform this analysis, we submitted our pyrosequencing dataset for annotation on the SEED platform [152-153]. Within SEED, metabolic pathways are classified into a hierarchical structure in which all of the genes required for a specific task are arranged into subsystems. At the highest level of organization, the subsystems include both catabolic and anabolic functions (for

example, DNA metabolism) and at the lowest levels the subsystems are specific pathways (for example, the synthesis pathway for thymidine). Using the blastx algorithm and an e-score cutoff of 0.001, the SEED database was able to assign ~60% of our sequences. The result of assigning these sequences to functional categories is shown in Figure 3.7. For comparison, we show in Figure 3.7 the mapping to functional categories from a recently published survey of 1,040,665 sequences from 45 microbial metagenomes collected from nine distinct biomes [151]. We note that when compared to the "average" profile of these nine biomes, the wastewater treatment plant presents a distinct metabolic signature. For example, compared to other biomes, the wastewater treatment plant contains nearly no genes coding for proteins involved in photosynthesis. We would expect this as the primary energy source for these microbes is the organic material being processed by the treatment plant. In addition, genes involved in the degradation of aromatic compounds are expressed at a much higher rate within the wastewater treatment plant than in other metagenomic systems. Again, we might expect this given the nature of household and industrial wastes present in sewage. Finally, we note that the Mallard Creek Wastewater Treatment Plant has no additional biological nutrient removal (BNR) facilities to treat phosphorus. Consistent with this, genes involved with phosphorus metabolism appear to be lower than the genes involved with nitrogen metabolism within the activated sludge (Figure 3.7).

3.5 Summary

We are at the beginning of a sequencing revolution. The 91 million base pairs of sequence data described in this paper were generated from a single sequencing run on a 454-FLX instrument generating over 6000 base pairs of sequence per dollar. This is

approximately a 10-fold lower cost per basepair than Sanger sequencing and moreover eliminates the costly and time-intensive step of creating a bacterial clone library. As new sequencing technologies continue to be developed, we can expect both the cost and the experimental effort associated with metagenomic sequencing projects to drop exponentially.

Perhaps the most surprising result in our study is the pronounced conservation of transposases across widely differing environments. While there is generally poor agreement between sequences from the Global Ocean Survey and known genomes [142], and between our wastewater genomes and known genomes (Figures 3.4-3.5), there are a few regions of conservation involving transposons (Figure 3.6) where there is a pronounced match between the metagenomes and the sequenced genomes. A substantial fraction of the contigs that could be assembled from our dataset involved strongly conserved transposases. It is an open question why transposons have escaped the pronounced sequence mutability that mark nearly all of the rest of bacterial genomes.

As in other metagenomic projects [10, 23, 142], our results point to the extraordinary diversity of microbial communities. Patterns of non-specific recruitment to known genomes suggest that even among the taxa that can be mapped to Genbank, the structure of much of the genomes of the most abundant organisms in the wastewater treatment plant is unknown (Figures 3.4-3.5). Despite the great diversity of microbes in the treatment plant, analysis at the protein level is surprisingly tractable with the sequences from the treatment plant displaying a distinct metabolic profile consistent with what we would expect from the plant's function (Figure 3.7). This suggests that despite the great complexity of microbial communities, next generation sequencing technology will be a

useful tool for monitoring changes in microbial processes across time and space. As treatment requirements become more stringent and monitoring expands to address a broadening group of compounds of concern, probe-free sequencing will accelerate the rate at which key microbial groups can be identified and selected for to optimize contaminant removal.

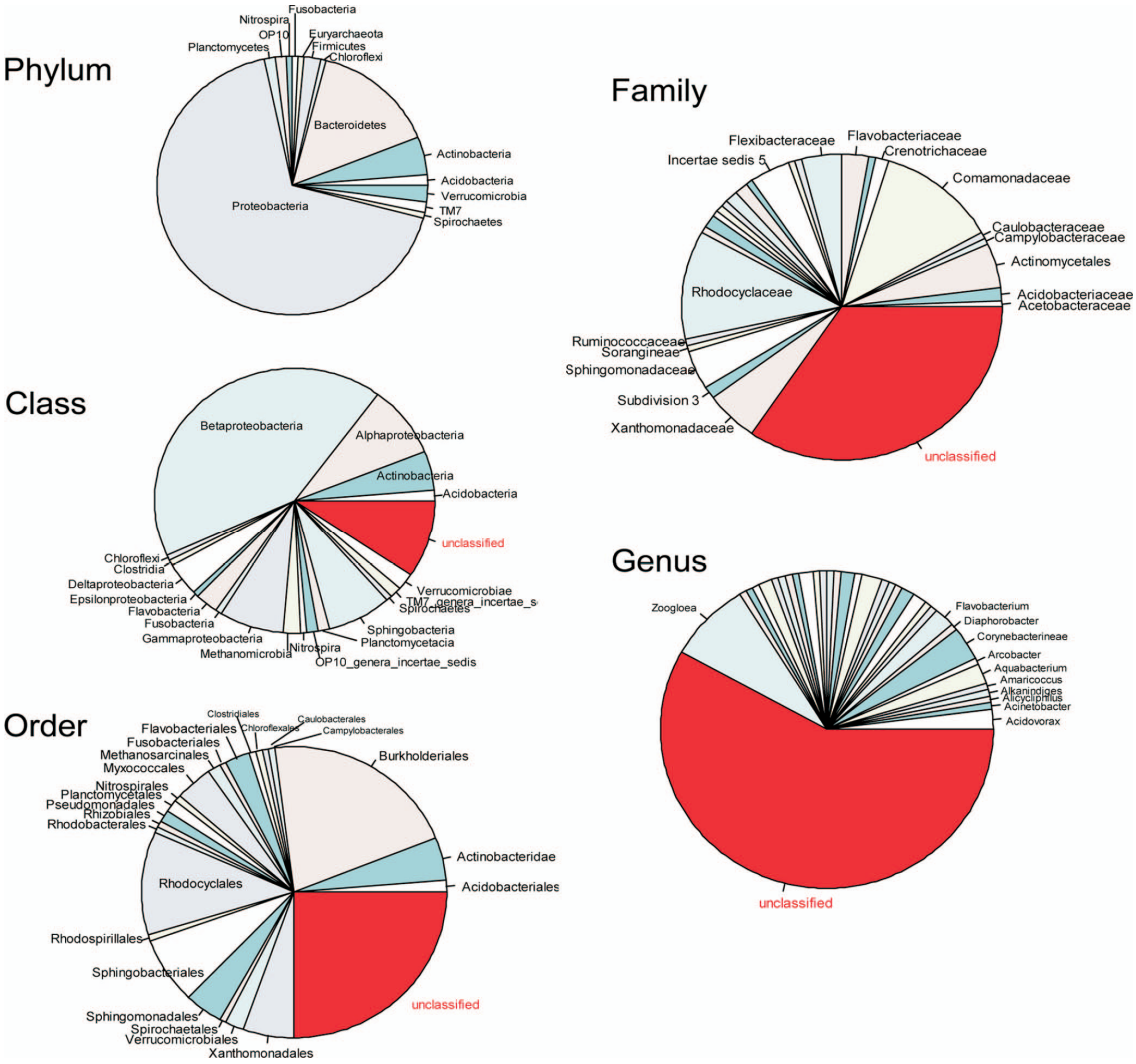


FIGURE 3.1: Pie charts show taxonomic assignments for 148 16S rRNA sequences within our dataset that could be classified to Phylum with an RDP confidence scores of >=80. At the phylum level, the Simpsons diversity index is 0.48.

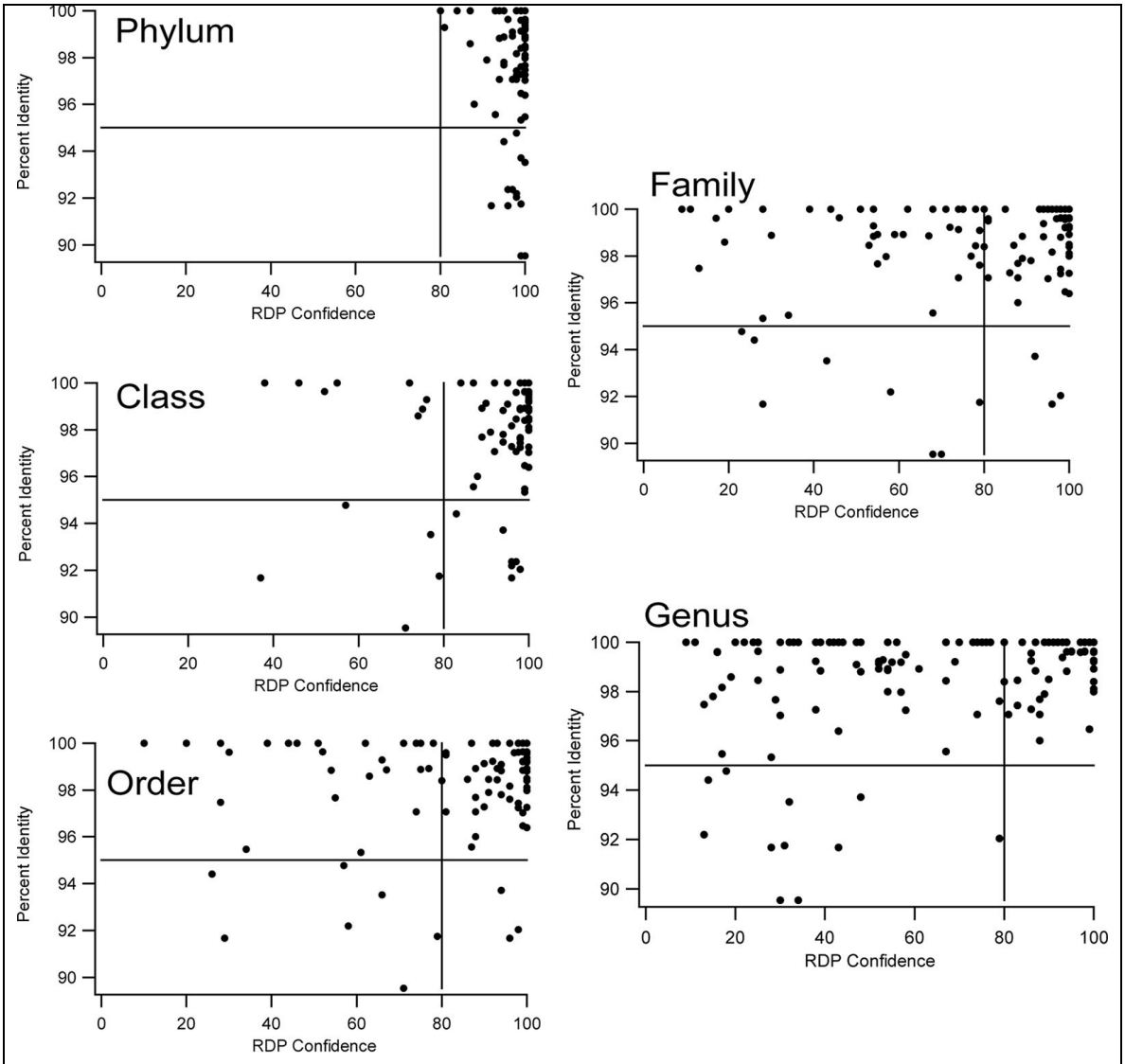


FIGURE 3.2: Results from the RDP classification algorithm for 148 16S rRNA sequences that can be assigned to the Phylum level with a confidence score of >=80. The x-axis of each graph shows the confidence in assignments as reported by the RDP classification algorithm. The y-axis of each graph shows the percent identity between our query sequence and the best BLASTN hit in the RDP database v 9.52. Horizontal and vertical lines indicate 95% sequence identity and an 80% RDP confidence scores.

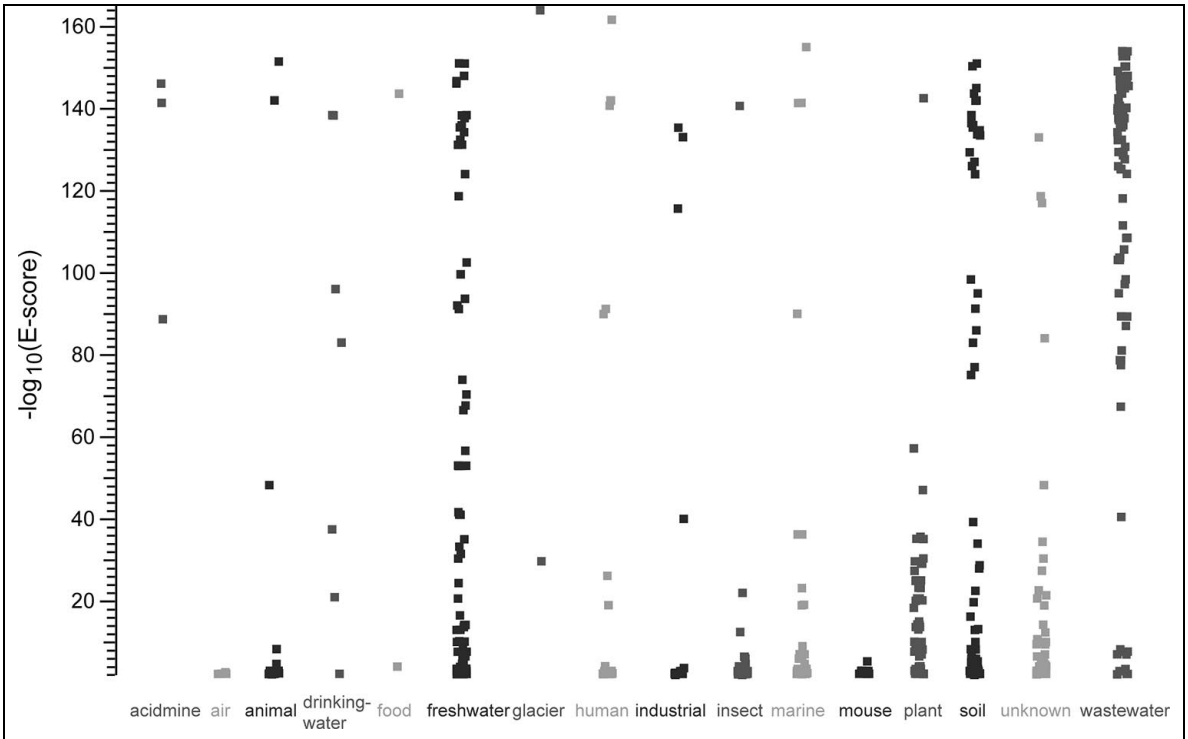


FIGURE 3.3: The location (as determined by manual annotation) and e-score of sequences from the 648 member pyrosequencing dataset that matched the 16S RDP database at an e-score cutoff of 0.01.

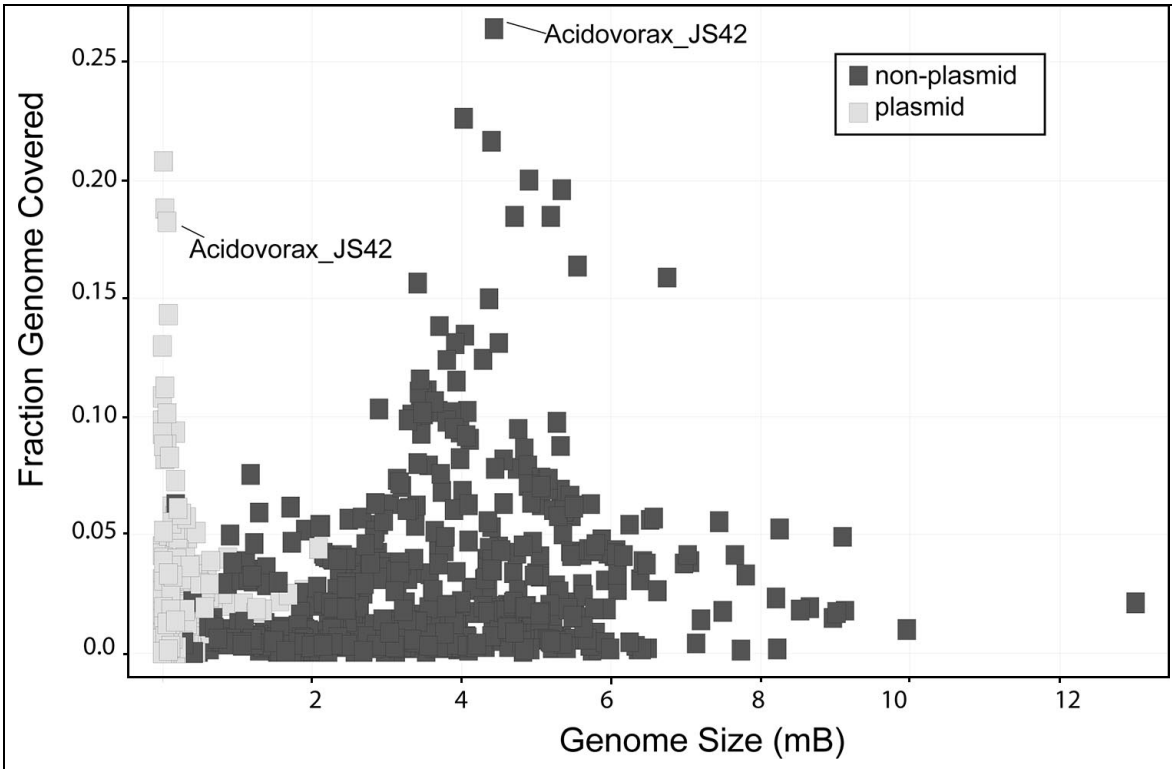


FIGURE 3.4: For each of the 1,442 assembled plasmids and chromosomes at NCBI, the fraction covered as a function of the size of each assembled sequence. Fraction covered is defined as the number of nucleotides in the assembled sequences that match at least one of our wastewater sequences divided by the total number of nucleotides in the assembled sequence.

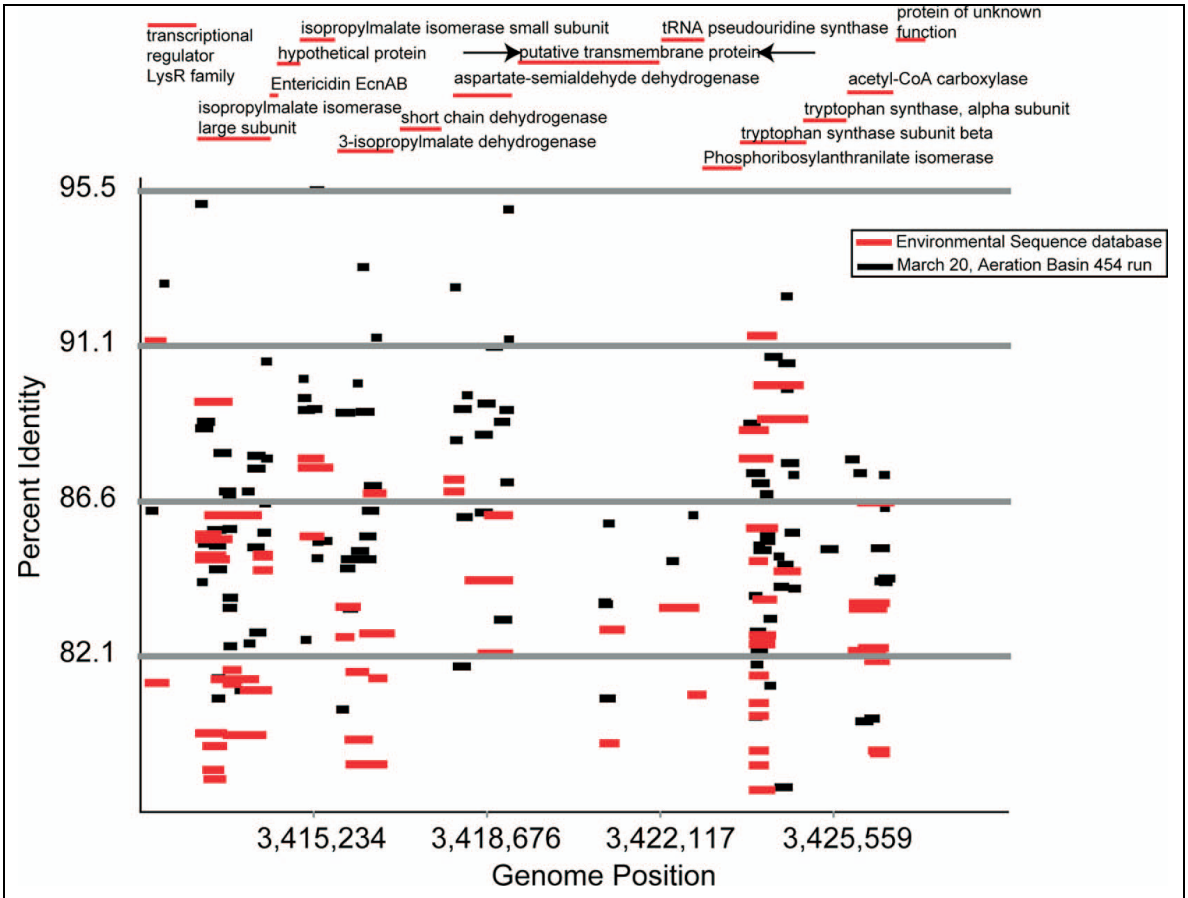


FIGURE 3.5: Non-specific recruitment against the Acidovorax sp. JS42 genome. Blast hits with alignment lengths below 75 nucleotides (for the March 20th run) or 250 nucleotides (for the Environmental Sequence database) were removed. Protein annotations are derived from the full NCBI core nucleotide report for the Acidovorax sp. JS42 genome. (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nuccore&id=121592436)

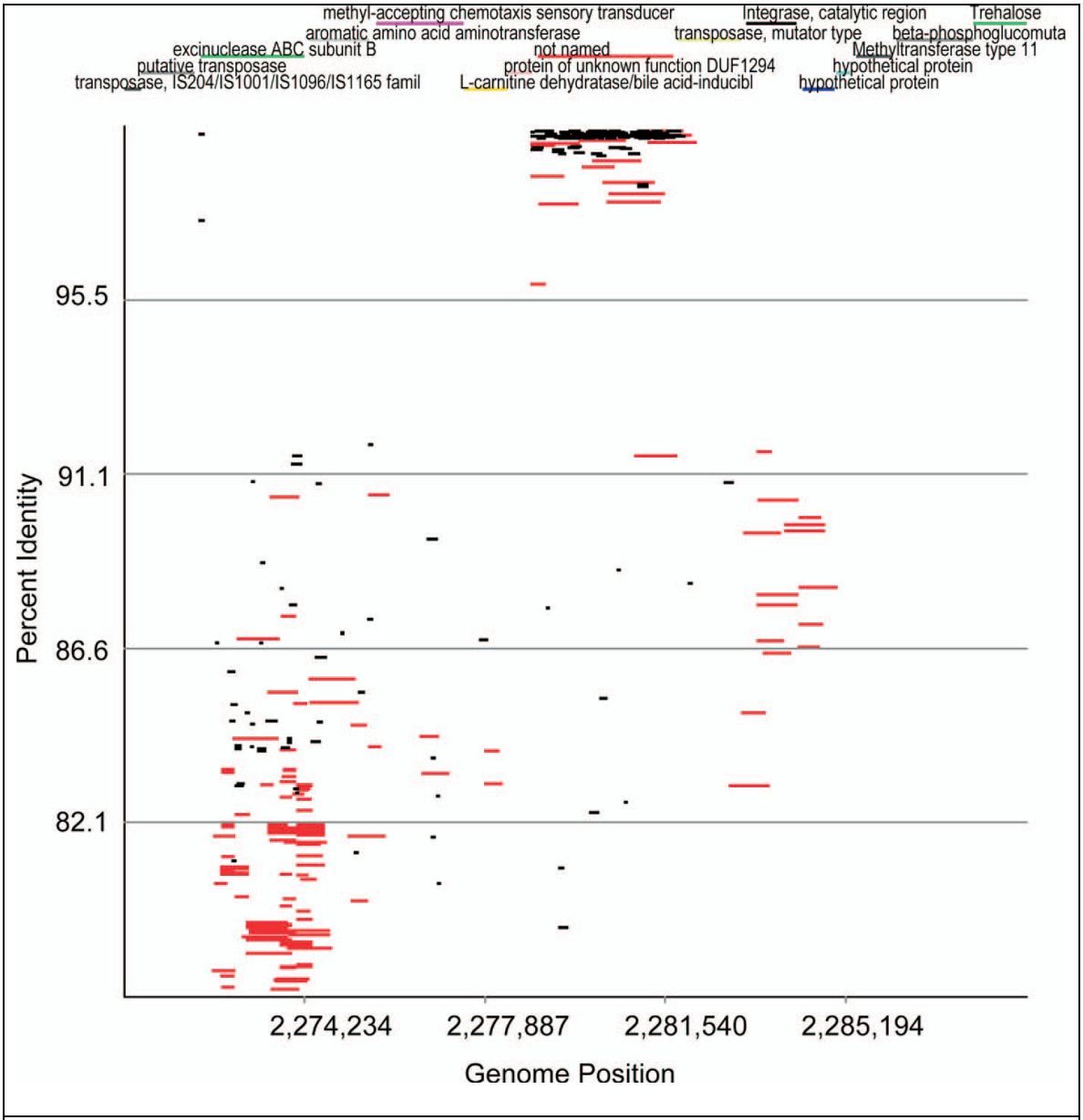


FIGURE 3.6: A region involving a transposase from the JS42 genome that shows an exception to the pattern of non-specific recruitment. For visualization purposes, a small amount of random noise has been added to the y-axis (as otherwise most of the hits to the transposase region would be superimposed). The sequences shown in red matching to the region of the transposase are from the Global Ocean Survey [142].

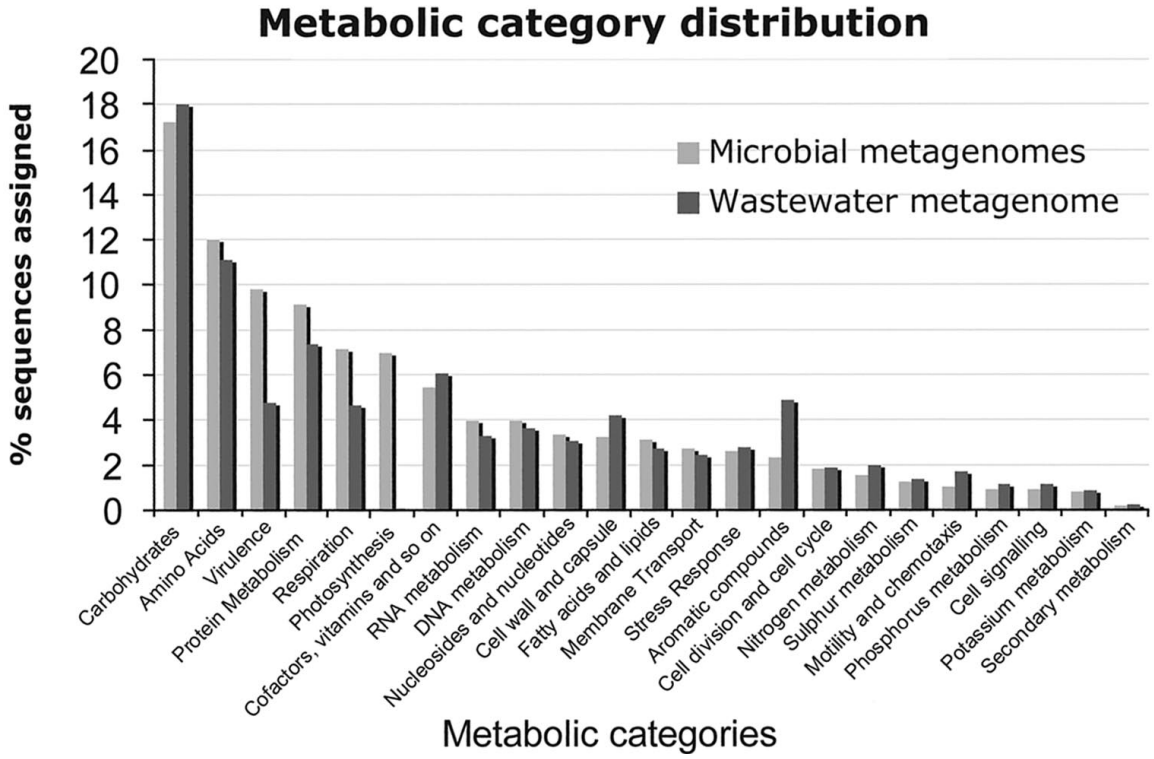


FIGURE 3.7: Functional categories provided for our dataset by the Seed server (http://www.theseed.org). The data for microbial genomes are averages from sequences gathered from multiple biomes [151].

TABLE 3.1: The top ten assembled microbial genomes as sorted by the number of hits recruited from our wastewater metagenome. The complete list of all assembled microbial genomes is given as Supplementary File 6(see methods).

numberHits	FractionGenome Covered	annotation
18110	0.26	gi|121592436|ref|NC_008782.1| Acidovorax sp. JS42,
17341	0.20	gi|120608714|ref|NC_008752.1| Acidovorax avenae subsp. citrulli AAC00-1
17100	0.16	gi|160895450|ref|NC_010002.1| Delftia acidovorans SPH-1
16800	0.20	gi|171056692|ref|NC_010524.1| Leptothrix cholodnii SP-6
15752	0.23	gi|124265193|ref|NC_008825.1| Methylibium petroleiphilum PM1
15695	0.22	gi|121602919|ref|NC_008781.1| Polaromonas naphthalenivorans
15468	0.18	gi|91785913|ref|NC_007948.1| Polaromonas sp. JS666
14735	0.16	gi|121607004|ref|NC_008786.1| Verminephrobacter eiseniae EF01-2
13590	0.18	gi|89898822|ref|NC_007908.1| Rhodoferax ferrireducens T118
11595	0.15	gi|119896292|ref|NC_008702.1| Azoarcus sp. BH72

CHAPTER 4: COMPARISON OF 16S rRNA GENE SEQUENCE BASED TAXONOMIC PROFILING TO WHOLE GENOME SEQUENCE BASED TAXONOMIC PROFILING METHODS

4.1 Abstract

One of the major steps in analyzing microbial communities is the estimation of the taxonomic composition of the community in question. 16S rRNA gene sequence based methods have been an accepted "gold standard" for taxonomic profiling of genomes as well as metagenomes. We evaluated methods that use whole genome sequences for determining the taxonomic composition of a community, to see if whole genome sequence based methods can replace 16S rRNA gene sequence based methods for taxonomic profiling. To achieve this, we compared methods that use sequences derived from PCR targeting the 16S rRNA genes of the community with whole genome sequences derived from shotgun sequencing of the community to see if they generate a similar or different taxonomic profile of the given community.

Not surpisingly, we find substantial differences between the two groups of methods with the degree of similarity decreasing from broad taxonomic levels (Phylum, Class) to more specific taxonomic levels (Family, Genus). At all levels of classification, however, there are assignments made by one group of methods but are missed by the other and vice-versa. This indicates that 16S rRNA gene sequence based and whole genome sequence based methods are complementary to each other, and that whole genome

sequence based methods cannot currently replace the 16S rRNA gene sequence based methods for taxonomic profiling of a community.

Amongst the whole genome sequence based methods evaluated, results show that the algorithm that only considers the 16S rRNA gene sequences within a whole-genome metagenomic dataset shows much better correspondence to PCR-derived 16S rRNA methods than algorithms that attempt to assign every read within a whole-genome dataset. Of the latter group of methods, BlastBestHit, MEGAN and WebCARMA, our results show that no method is obviously superior to any other. Also, the fact that different methods report different diversity indices for the same community proves that the method selected for taxonomic profiling determines the depth of the taxonomic composition extracted from the community.

4.2 Background and significance

The choice of taxonomic profiling methods used for community analysis differ based on whether the 16S rRNA gene sequences (targeted by PCR) are available or whole genome sequence reads (generated by random shotgun sequencing) are available. Some methods use 16S rRNA gene sequences, generated from targeted PCR or mined from random whole genome datasets, to describe the taxonomic composition of the community in question. Other methods (such as BlastBestHit, MEGAN [154] and CARMA [155]) place random whole genome reads into a taxonomic framework based on their similarity to one or more reference databases.

For almost three decades, 16S rRNA gene based taxonomic profiling has been the classical "gold standard" approach to assess the microbial composition in environmental gene surveys [34, 156]. 16S rRNA gene sequences are present in all bacteria and consist

of multiple conserved regions interspersed by variable regions. Conserved regions can be used to design primers while the variable regions can be used for phylogenetic assignment. These features make the 16S rRNA gene a good candidate for phylogenetic analysis by PCR[157]. Despite being a powerful tool for phylogenetic and taxonomic analysis, the 16S rRNA gene sequence based analyses have several limitations. There is significant variation in the number of copies of the 16S rRNA gene present in the genome of different species. This variation can complicate quantitative estimates such as relative abundance of a particular sequence type (phylotype) in an environment [158]. Another limitation of this method is that in the PCR amplification step of sequencing, the broadly targeted "universal" PCR primers designed to amplify all members of a major taxonomic groups (e.g., all bacteria, or all archaea) are not in-fact "universal". This is because all members of a taxonomic group do not share identical sequences even in the conserved region; primer bias, therefore, remains a problem [159]. Even the best designed primer pairs tend to be biased towards some evolutionary group over others making the resulting taxonomic profiles generated by this method not a true representation of the community. A third drawback of 16S rRNA gene based method in general is that the taxonomic assignment made using this method is not always accurate and does not necessarily reflect the true phylogeny of the organism [160]. This inaccuracy can be attributed to many factors including biased databases, lateral gene transfer and different rates of evolution for different genes within a microbial genome. Whole genome sequence based taxonomic profiling methods promise to overcome some of the above mentioned limitations, but these approaches generate their own set of problems. Typically, when whole genome sequence reads from metagenomic studies are used for taxonomic

profiling of a community, phylogenetic categories are assigned to the randomly generated sequence reads based on their homology to known genes in sequence databases. One approach is to find the 16S rRNA gene reads from within the whole genome sequence set and use those 16S rRNA reads for taxonomic profiling of the community[44]. One drawback of this approach is that this method relies on an adequate number of 16S rRNA gene sequences being present in the metagenomic sequences. Usually the number of reads containing stretches of 16S rRNA gene sequence, long enough to be used for this purpose, are very small (for example less than 500 out of ~300,000 in the wastewater dataset[44],[117] and less than 1000 out of ~500,000 in the human gut microbiome dataset).

An alternative, to the use of just the 16S rRNA gene sequences from whole genome datasets, is to use all of the sequence reads (or assembled contigs) as input for a search against a protein or nucleotide database to find homologues whose taxonomy can then be assigned to the reads. One of the simplest examples of this approach is the “BlastBestHit” method, which assigns taxonomy to a sequence using the taxonomy of the topmost BLAST match. The shortcomings of BLAST based methods include the requirement of a sufficient sequence length and the existence of close homologues in the reference database. Another BLAST-based analysis is MEGAN[154], which takes the search results of the sequence reads (using BLAST or any sequence comparison tool) against any protein or nucleotide as its input and assigns taxonomy to reads based on the lowest common ancestor (LCA) of the selected hits. For instance if a read has 10 hits at the genus level, which meet the similarity threshold, and if >2/3 of the hits have the same taxonomy at the genus level then the read gets assigned to that particular genus. If there is

no consensus at the genus level, the algorithm looks for a consensus at the family level and so on, up the tree, until it finds a consensus at a particular taxonomic level. The read then gets assigned to the consensus taxon at that taxonomic level. Due to this strategy, all sequences, irrespective of their conservation level, are assigned taxonomy by MEGAN; species specific sequences are assigned to corresponding species and highly conserved sequences are assigned at higher taxonomic levels. So the taxonomic level assigned to a sequence also implicitly indicates the conservation level of that sequence-read. The consensus finding step of the algorithm helps avoid making erroneous assignments due to horizontal gene transfer and database bias by not making any assignment at that taxonomic level at all (due to lack of consensus if a sequence has matches to many different taxa). The biggest shortcoming of the MEGAN method is that its underlying algorithm only takes into account presence or absence of matches for the reads at the given score threshold. Once the matches are found at a given sequence similarity threshold, the matches are not ranked by their level of similarity, instead they are just weighted equally while assigning taxa.

Homology based approaches for assigning taxonomy to sequence reads were further extended to the protein level and one notable application is the tool CARMA[155], which starts out by looking for Pfam domains within the sequence-reads. Once it finds matches it builds phylogenetic trees from the sequence-reads and their reference sequences and then classifies them. The CARMA algorithm has two distinct modules; the first step identifies Pfam domains or fragments in the unassembled reads using the Pfam profile hidden Markov models (pHMMs). In this step the biggest advantage of the CARMA method, its ability to assign taxonomy to short sequence fragments, generated by “next-

generation" sequencing methods (454, Illumina etc.), is harnessed. This ability is conferred to this method because pHMMs are very efficient in detecting short conserved functional sequences within the sequence-reads. The Pfam domain and protein family matches identified within the reads of the metagenomic sample are called "environmental gene tags" (EGTs), which can be used in the next step for quantitatively characterizing the metagenome. In the second step of this method, a phylogenetic tree is constructed for each EGT (environmental gene tag) with its matching Pfam family and the environmental gene tags are classified based on their phylogenetic relationships (proximity in the tree) to the Pfam family members with known taxonomies. CARMA is computationally demanding but has been shown to exhibit high accuracy for a wide range of taxonomic groups; sequence-fragments as short as 80 bp and EGTs as short as 27 amino acids have been phylogenetically classified up to the rank of genus [155]. The Pfam domain assignments made by CARMA not only help in taxonomic profiling but also in functional profiling of the metagenome. This group has recently released the web-server version of their method, WebCARMA[161]. The major drawback of the WebCARMA method is that there is an upload limit of 100MB per month per user, so large metagenomic datasets cannot be processed using this method.

For classification of the sequences generated by whole genome shotgun sequencing of a metagenome, a diverse range of methods mentioned above, have been used with dramatic differences in classification results, depending on both underlying algorithms and parameters[162]. In this dissertation chapter, we compare the results of these different methods on datasets for which we have both PCR-based 16S rRNA sequences

and whole genome-metagenome datasets to ask if taxonomies constructed with these two methods are identical.

4.3 Materials and Methods

There are many taxonomic profiling methods available to assess taxonomic composition of a given community [163], [46], [162], [164], [165], [154], [155] but due to time and resource constraints we limited our comparison to a few of these methods. The various taxonomic diversity assessment tools, compared in this chapter, were either downloaded and run according to the authors' instructions or re-implemented in Java whenever the source codes were unavailable (see computational methods below). The analysis path followed is shown in the flowchart described in Figure 4.1.

Two metagenomic datasets, the wastewater dataset from our lab, an environmental metagenome, described in chapter 3 of this dissertation above, and a human gut microbiome dataset from one of the subjects, TS19, of the 31 monozygotic twin pairs and 23 dizygotic (DZ) twin pairs, in the Twin study[52] performed by Jeff Gordon's group, were chosen for analysis. We chose these two datasets not only because both these datasets had the whole-genome and 16S sequence libraries available but because they differ in their compositional complexity and their representation in sequence databases. For both datasets, the PCR generated 16S rRNA gene sequences were submitted to the RDP classifier algorithm[46] to get taxonomic assignments for the sequences at a confidence threshold of 80%. The whole genome sequences, from both datasets, were submitted to the various whole genome sequence classification methods, 16sMined, BlastBestHit, MEGAN and to WebCARMA to get the respective classifications for the reads.

4.3.1 Computational Methods

4.3.1.1 Targeted 16S rRNA gene (PCR generated) based taxonomic profiling

The 16S rRNA gene sequences, which have been targeted by performing PCR on DNA extracted from the community in question, are the input sequences in this method. The most commonly used primer pairs target the V1-V2, V6, V6-V7 or V2, variable regions in the 16S rRNA gene. These sequences, V1-V2, V6 and V6-V7 in the wastewater dataset and V2 and V6 in the twin-study dataset, were submitted to the RDP classifier 2.0, a Naïve Bayesian Classifier and the RDP classifications are assigned to the 16S rRNA gene sequences. The taxonomic assignments made to the sequences at every taxonomic level, with a RDP confidence threshold of >=80% were only considered.

4.3.1.2 16sMined

A BLAST database was created from all available 16S rRNA gene sequences (current_prokMSA_unaligned.fasta) downloaded from the Greengenes database (a 16S rRNA gene sequence database)[166]. For each query sequence, from the metagenomic datasets, we performed a BLASTN search against this database at a stringent e-Value cut-off of $e = 10^{-8}$. Query sequences, which found matches in the 16S rRNA gene sequence database at this low e-Value threshold, were considered as probable 16S rRNA gene sequences. These mined 16S rRNA gene sequences were submitted to the RDP classifier algorithm for classification. The taxonomic assignments made to the sequences at every taxonomic level, with a RDP confidence threshold of >=80% were only considered.

4.3.1.3 16sMerged

Hamp et al[44] suggested that different 16S rRNA gene regions capture different fractions of the community composition due to primer bias. Therefore, for our

comparative analyses, we merged (pooled) the taxonomic profiles interpreted by the 16S rRNA gene sequences derived from various primer pairs by combining the assignments made by all the 16S sequences available for the dataset.We merged V1-V2, V6-V7 and V6 for the wastewater dataset; and V2 and V6 sequences for the Gordon twin-study dataset. Other methods (averaging the profiles and normalizing the profiles to a fixed number of sequences and then averaging them etc.) of merging the taxonomic profiles of the 16S rRNA regions for a given dataset were tried but all gave very similar results so the method described above was used.

4.3.1.4 BlastBestHit method

This method was used for classification of reads generated by whole genome shotgun sequencing of the 2 communities chosen in this study. The sequences were searched against a database of all sequenced bacterial genomes (nucleotide) using BLASTN at a relaxed e-Value threshold of 0.01. The taxonomy of the top hit for each sequence is the taxonomy assigned to it. Sequences which have no hits at this e-Value cut-off remain unclassified.

4.3.1.5 MEGAN

This program is written in JAVA (http://www.java.com/en/); the source code for this algorithm has not been released but the JAR (Java Archive) files are available for download at http://www-ab.informatik.uni-tuebingen.de/software/megan. In the first step of the analysis, a database was created from all sequenced bacterial genomes (nucleotide). The DNA reads, from our datasets (wastewater dataset and twin-study dataset) were compared against this database using a BLASTN search. The BLASTN

search results are the input for MEGAN, which uses its LCA strategy to assign taxonomy to the reads.

4.3.1.6 WebCARMA

We used the web application version of the CARMA algorithm, WebCarma[161] for classification of our reads. The upload limit is 100MB, so we preprocessed the sequences in our Wastewater dataset to remove duplicates and filter out sequences with lengths lesser that 75bp and greater than 280bp. The filtered version of the dataset was used for comparison across all the algorithms. The human gut microbiome whole genome dataset was less than 100MB so we uploaded it, as is, to the WebCarma portal for taxonomic assignment.

4.3.2 Comparative Analysis

4.3.2.1 NCBI namespace to RDP namespace

The taxonomic assignments for the whole genome sequence reads were converted from the NCBI naming format to the RDP namespace. To achieve this, all the sequenced bacterial genomes were downloaded from the NCBI genome database. 16S rRNA gene sequences from these genomes were extracted and submitted to the RDP classifier 2.0 for RDP classification of all the bacterial 16S rRNA gene sequences. Most bacteria have more than one 16S rRNA gene sequence; in those cases the consensus taxonomy of all the 16S rRNA genes within a given bacterium was chosen as the RDP taxonomy for that bacterium. For instance if a bacterium has five 16S sequences, the taxonomic classification of more than 50% of these 16S sequences at the genus level is the consensus taxonomy of that bacterium.

Once all the reads from the PCR generated 16S rRNA gene datasets and the whole-genome sequence datasets were given a taxonomic assignment and the final taxonomic profiles were converted to the RDP namespace, comparison between all the methods was done. At each taxonomic level, starting from the Phylum level to the genus level, taxonomic assignments were counted and a pseudocount of 1 was added to all taxa counts before log transformation. To assess the similarity between any two methods, comparisons were performed, by creating scatterplots, on the log transformed counts at all taxonomic levels. Since the relationship between taxonomic assignment methods did not meet the assumption of linearity, Spearman Rank Correlation (rho) was used to measure the agreement between various methods. Shannon Diversity indices of each of the datasets, using the various methods, were also compared.

4.3.3 Statistical methods

Non-parametric correlations (Spearman Rank correlations) were generated using JMP (SAS Institute). Shannon-Wiener Diversity indices, H, were calculated using the equation, $H = -\sum Pi\ (lnPi)$, where Pi is the proportion of each taxon within the method.

4.4 Results

4.4.1 16S rRNA mined method is more similar to the PCR targeted 16S rRNA methods than the whole genome sequence based methods.

As an initial step of our analysis, we compared the 16S sequences mined from the whole genome sequences from the wastewater dataset to the PCR targeted 16S sequences from the same environment. Just as reported by Hamp et al[44], our results indicate that the mined 16S rRNA gene sequences are similar to both the whole genome sequences as well as the targeted 16S rRNA gene sequences (Figures 4.2a and 4.2b) in their ability to

assess the community's taxonomic profile. The mined method is comparatively more similar to the PCR targeted 16S rRNA methods than the whole genome sequence based methods and in both our datasets, the similarity decreases as we go from the Phylum level to the Genus level, dropping to lower levels of agreement or no agreement at the Genus level (Figures 4.2a and 4.2b). The 16sMerged method correlated better, with the 16sMined method, than the individual 16S rRNA methods. Since the 16sMerged method serves as a combined representative of the various 16S regions, it was used as a proxy for all the 16S regions, in comparisons of 16S rRNA gene sequence based methods to the whole genome sequence based methods (Figures 4.5a and 4.5b).

Also, as shown by Hamp et al[44] and by our own results (Figures 4.3a and 4.3b,4.4a and 4.4b), the agreement at Phylum level is being driven by "abundant" taxa and the differences at the genus level are due to the non-abundant taxa.

4.4.2 The two groups of methods (16S and WGS) agree at broader taxonomic levels but the degree of correlation decreases towards the specific taxonomic levels.

Comparison between PCR 16S rRNA gene sequence based taxonomic profiles and whole genome sequence based taxonomic profiles, of the wastewater dataset (Appendix B, Supplementary Figures 1a and 2a; Appendix B, Supplementary Table 1) and the human gut microbiome dataset (Appendix B, Supplementary Figures 1b and 2b, Appendix B, Supplementary Table 2), showed that there is a good degree of correlation between PCR 16S rRNA methods and whole genome methods at the phylum level but the correlation decreases considerably as we go down to the genus level (Figures 4.5a and 4.5b). Within the PCR 16S rRNA gene sequence methods, the V1-V2 region in the wastewater dataset and the V2 region in the Gordon dataset performed marginally better

than the other primer pairs and than the 16sMerged method (Appendix B, Supplementary Tables 1 and 2).

4.4.3 16sMined method is the only whole genome sequence based method that shows potential for replacing the PCR targeted 16S sequence based methods.

As shown by our results (Figures 4.5a and 4.5b; Appendix B, Supplementary Tables 1 and 2), the 16sMined method is the best match to the PCR targeted 16S gene sequence based methods at all levels of classification and it is the only whole genome sequence method that shows potential for replacing the PCR targeted 16S sequence based methods. The advantage of the 16sMined method is that it overcomes primer bias; one of the major drawbacks of the PCR targeted 16S methods. However, the biggest shortcoming of this method, as discussed earlier, is that very few 16S rRNA gene sequences are produced by this method to give a reasonable enough taxonomic profile of the community. Therefore, this method of taxonomic profiling of the community cannot currenly replace the PCR 16S rRNA gene sequence based methods but may become viable as the decreasing cost of sequencing in the future increases the size of shotgun whole-genome datasets (with possibly more 16S rRNA sequences to extract).

4.4.4 Performance of the Whole Genome Sequence based methods is driven not only the by underlying algorithm but also by the community complexity and by the database bias.

When we compared the 16sMerged method (a combination of taxonomies generated by the PCR targeted 16S regions available for a given dataset) to the three whole genome sequence (WGS) based methods with different underlying algorithms (BlastBestHit, MEGAN and WebCARMA), our results show that these methods are very similar, to each other, in their perfomance (Figures 4.5a and 4.5b; Appendix B, Supplementary table

1 and 2). For the wastewater dataset, all 3 methods were nearly identical (Figure 4.5a), whereas for the human gut microbiome dataset (Figure 4.5b), WebCARMA slightly outperforms the other two methods at all taxonomic levels except the Phylum level. As expected, BlastBestHit and MEGAN, where the underlying search algorithm (BLAST) and the database (sequenced genomes) are the same, seem to be more similar to each other than to WebCARMA. Also all the WGS methods were slightly better correlated with the PCR 16S Method (16sMerged) in the human microbiome dataset than in the wastewater dataset (environmental metagenome). The human microbiome is by far less taxonomically complex than the wastewater community and also has been studied more so has a better representation in the sequence databases (both protein and nucleotide). This indicates that in addition to factors such as the underlying algorithm, the complexity of the community sampled and the biases in the database searched play a significant role in the performance of the taxonomy profiling method.

4.4.5 Different methods produce different profiles of the same community as shown by Shannon Diversity measurements.

Shannon diversity indices for a given environment (Wastewater community or Human gut microbial community) are considerably different using different methods (Figures 4.6a and 4.6b). As discussed in the previous section, this discrepancy could be not only due to obvious reasons such as the efficacy of the algorithm but also due to the not so obvious reasons such as possible database associated bias. The database bias is introduced because different methods rely on different publicly available databases to confer taxonomic assignments to the reads.

4.5 Discussion

With the range of methods available for evaluating the taxonomic composition of a community, the choice of the right method to use can get extremely confusing. A systematic comparison of the methods used for determining the taxonomic composition of a metagenomic community is needed, and one of the goals of this study is to fill that void. Other studies have done comparisons between taxonomic profiling methods but most of these studies were partial comparisons[167] with the introduction of new or improved algorithms[168]. None of these studies involve a thorough evaluation of the types of taxonomic profiling algorithms used but are limited to comparison of the new methods developed by that particular group to earlier taxonomic profiling algorithms[162]. A recent study by Hamp et al compared the targeted 16S rRNA sequences to the randomly generated 16S rRNA sequences (mined from the metagenome), in terms of their ability to generate the taxonomic profile of the community, and showed that profiles generated by either method were in general agreement but this agreement did not extend to the rare taxa[44]. Their study demonstrated that the choice of primers targeting the 16S rRNA gene can also have an effect on the taxonomic profile reported and our results attest to that. Our study, unlike its predecessors closely evaluates both PCR 16S rRNA gene sequence and whole genome sequence based taxonomic methods with the purpose of providing comparative data and results that can help researchers make an informed choice.

The advantage of the whole genome sequencing approach is due to the fact that it samples of the entire DNA present in a community but this method works best for identifying the abundant organisms in a community. On the other hand, there is the rRNA-PCR method and even though it targets only a single gene, it allows the

characterization of the less abundant organisms in a community. Metagenomic sequencing does produce 16S rRNA gene sequences which can be analyzed in same way as PCR-generated 16S rRNA gene sequences. Our results indicate that profiling the 16S rRNA gene sequences extracted (mined) from the community metagenome serves as an important cross check for both approaches and that it is the only whole genome sequence method that closely matches the PCR targeted 16S rRNA gene sequence methods. While this definitely should be the preferred method of taxonomic assessment from whole genome sequence datasets, it is not presently a replacement for the PCR targeted 16S rRNA gene sequence methods. Whether the 16S rRNA sequences originate as a result of targeting the 16S rRNA gene or are a result of whole genome sequencing of environmental DNA (metagenomic sequencing), one of the greatest incentives for concentrating on 16S rRNA gene sequence based studies is the ever-expanding database of 16S rRNA gene sequences from cultured organisms and environmental samples being deposited regularly.

Some researchers view whole-genome sequence based taxonomic methods as a replacement for rRNA gene PCR based methods [167]. The results of our systematic comparison indicate that these methods are not mutually exclusive (with very different outcomes) but have some degree of overlap and the degree of overlap decreases from broader taxonomic levels (Phylum, Class) to narrower taxonomic level (Family, Genus). At every taxonomic level, assignments made by one method are missed by the other and vice-versa. Since these are complementary approaches, it is obvious that whole genome shotgun sequencing cannot currenly replace 16S rRNA gene based taxonomic profiling for assessing the community composition.

Of the three whole genome sequence based methods (BlastBestHit, MEGAN and WebCARMA), with different underlying algorithms, we consider that all of them are very similar in terms of their agreement with the PCR 16S sequence based methods. The efficiency of these methods, as with other taxonomic profiling methods, is dependent not only on the algorithm but also on the complexity of the community being profiled and on the database being searched. Finally, since the taxonomic profile of the same community, as reported by the various methods (see Figures 4.6a and 4.6b) is different, the choice of the method used in taxonomic profiling of a community does indeed have an effect on the taxonomic profile assessed. The factors affecting the performance of the various methods are possibly the same ones that offer an explanation as to why all the methods give different Shannon Diversity indices when the same community is profiled.

The value of this study lies in the fact that it illustrates that when the true composition of an environment is unknown, it is better to use as many of the complementary methods as possible to get the best picture of "who is there". Our results also reiterate, what other such studies [44], [162] have found, that the various taxonomic assessment methods are only "snapshot" tools because they only capture some portion (not the entire range) of the community's diversity. Since each of the methods evaluated in this study differ not only in their range but the in their depth, we conclude that the choice of the taxonomic profiling method/methods should depend on the level of resolution desired from the community under study.

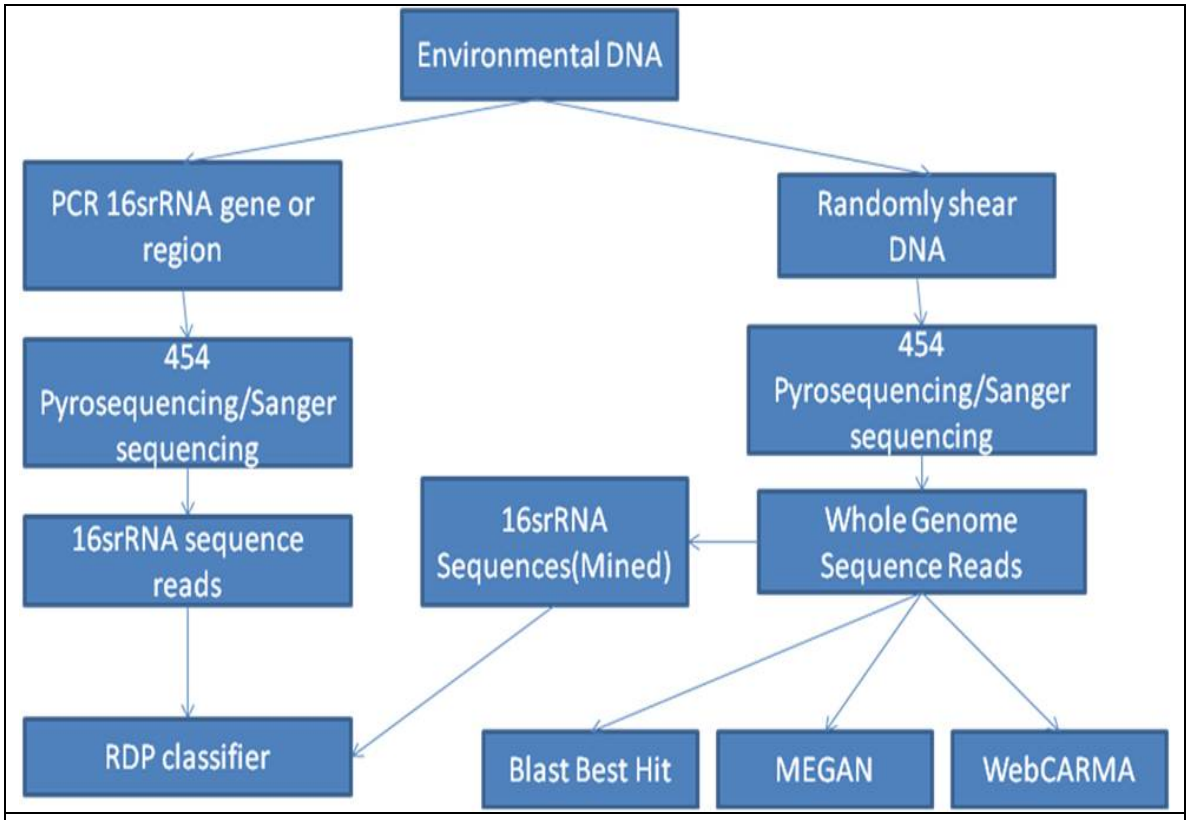


FIGURE 4.1: Shows a flowchart describing the analysis path followed in comparison of taxonomic profiling methods.

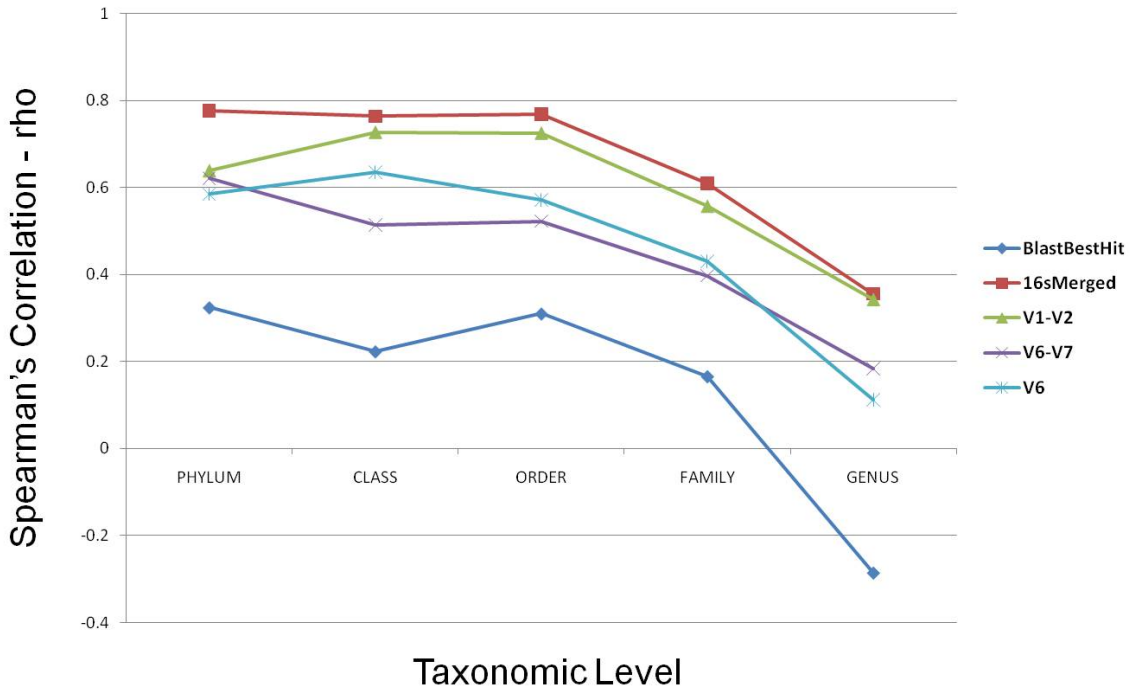


FIGURE 4.2a: Comparison of the 16S sequences mined from the wastewater metagenome to the PCR targeted 16S sequences and whole genome sequences from the same environment.

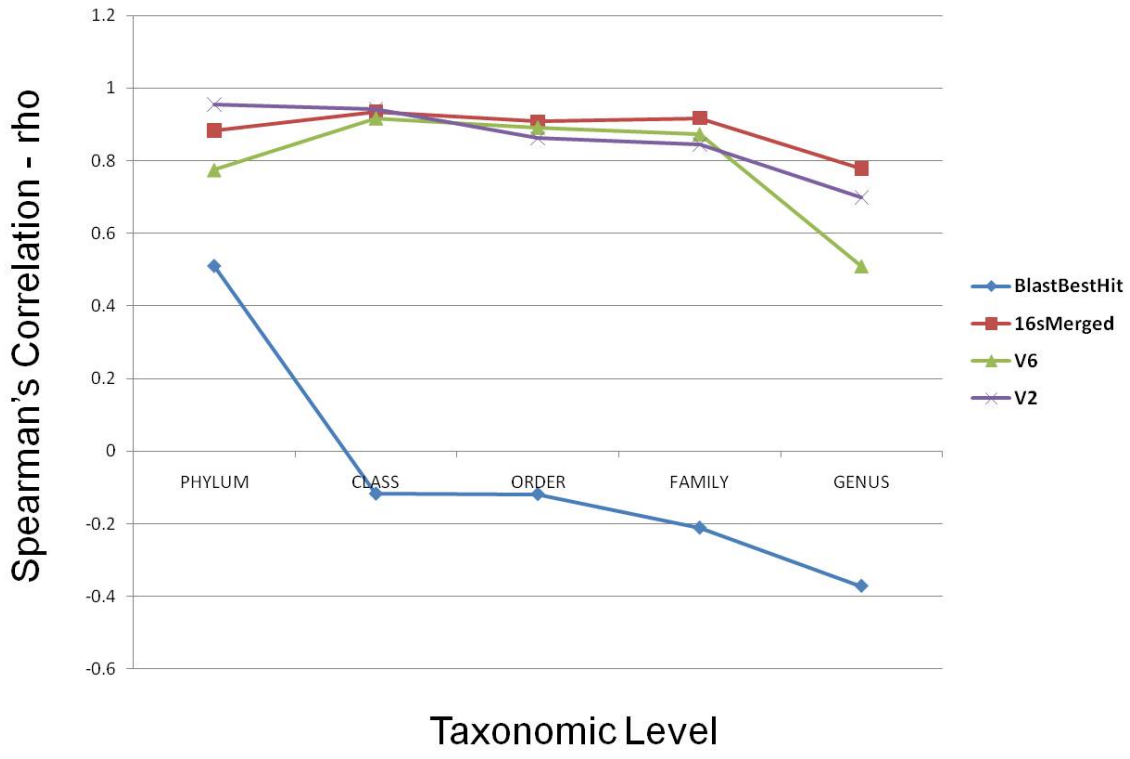


FIGURE 4.2b: Comparison of the 16S sequences extracted (mined) from the metagenome of the human gut microbiome to the PCR targeted 16S sequences and whole genome sequences from the same environment.

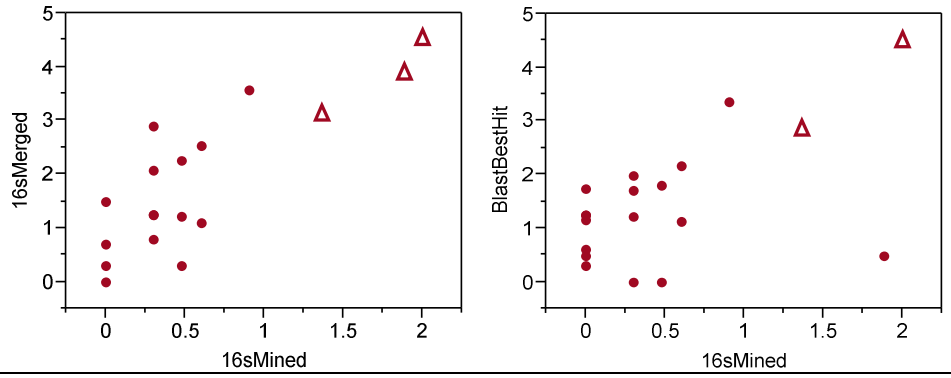


FIGURE 4.3a: Wastewater dataset: Scatter-plots showing the high level of agreement between 16sMined methods and the16sMerged method and both BlastBestHit method at the Phylum level. The abundant taxa driving the agreement are marked by red triangles.

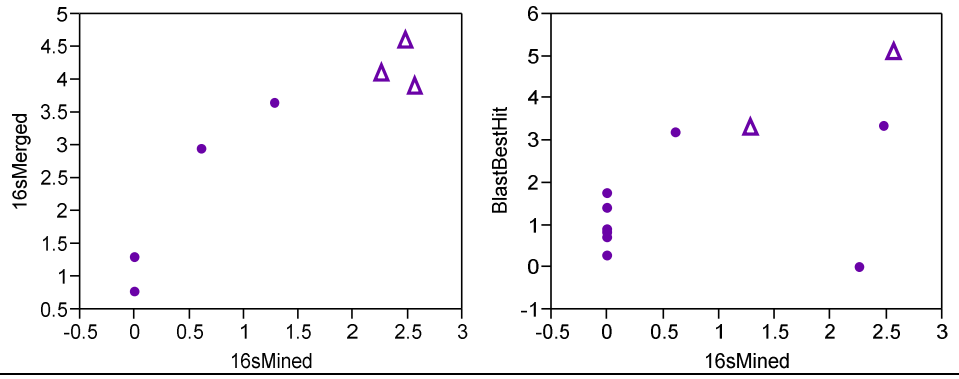


FIGURE 4.3b: Human gut microbiome dataset: Scatter-plots showing the high level of agreement between 16sMined methods and the 16sMerged method and both BlastBestHit method at the Phylum level. The abundant taxa driving the agreement are marked by purple triangles.

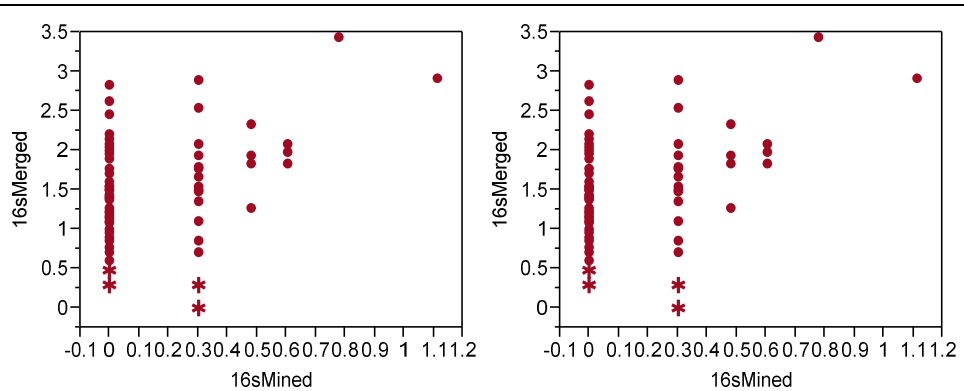


FIGURE 4.4a: Wastewater Dataset: Scatter-plots showing the relatively lower level of agreement between 16sMined methods and 16sMerged method and BlastBestHit methods at the Genus level compared to the Phylum level. Some of the low abundance taxa responsible for the lower level of agreement are marked with red asterisks.

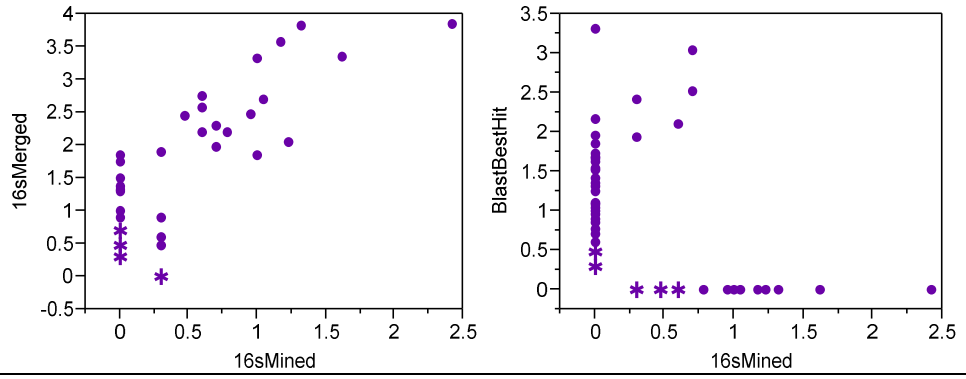


FIGURE 4.4b: Human gut microbiome dataset: Scatter-plots showing the relatively lower level of agreement between 16sMined methods and 16sMerged method and BlastBestHit methods at the Genus level compared to the Phylum level. Some of the low abundance taxa responsible for the lower level of agreement are marked with purple asterisks.

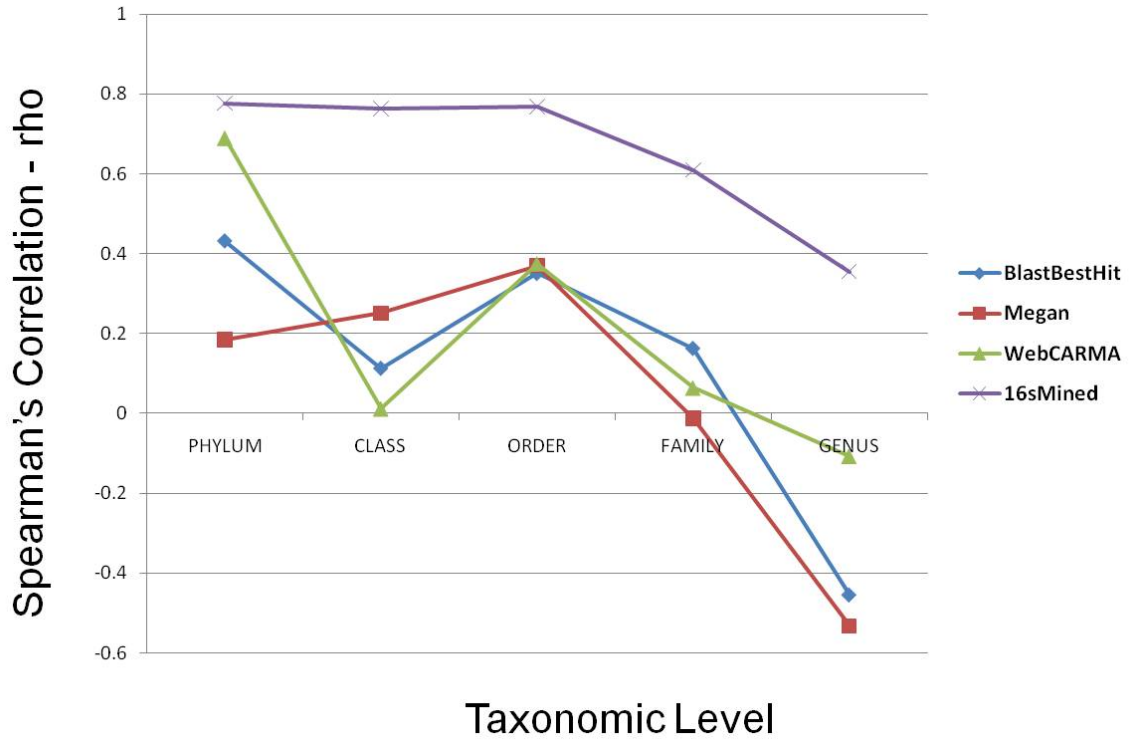


FIGURE 4.5a: Comparison of the 16sMerged method(PCR 16S method) to all the whole genome sequence based methods shows that, for the wastewater metagenome, the 16sMined method (WGS 16S method) performs best amongst the 4 whole genome sequence based methods evaluated and the other 3 methods are very similar to one another.

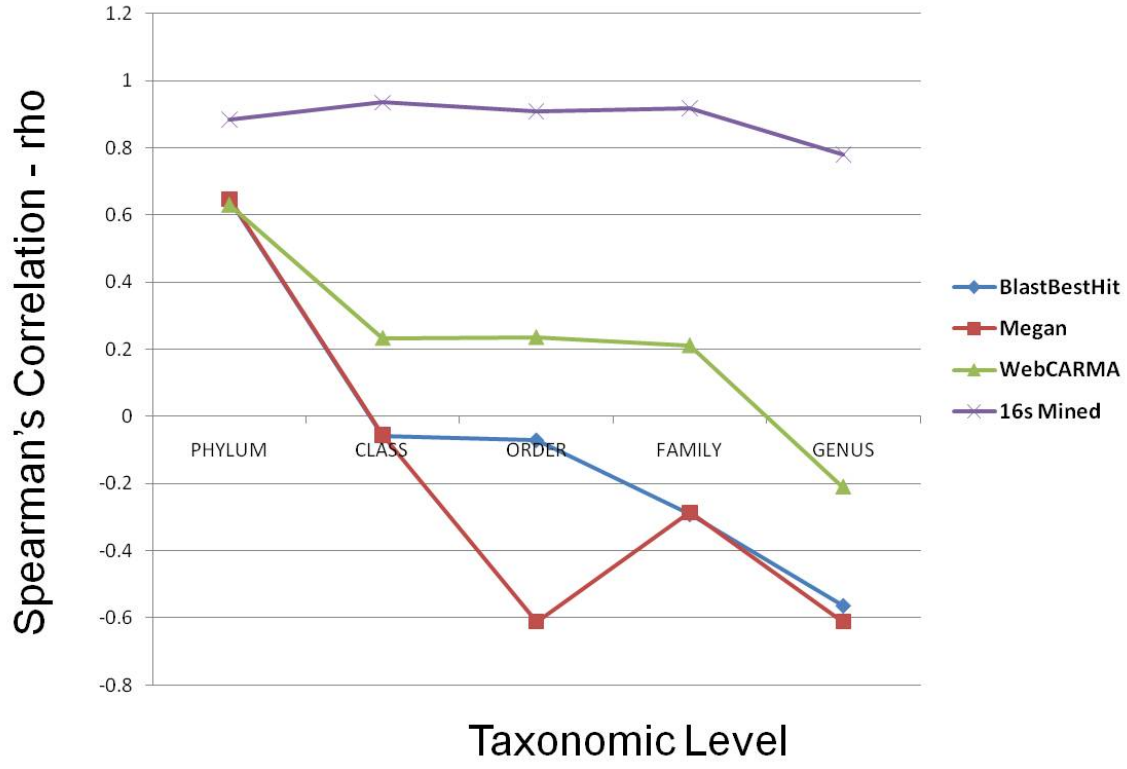


FIGURE 4.5b: Comparison of the 16sMerged method (PCR 16S method) to all the whole genome sequence based methods shows that, for the human gut microbiome dataset, the 16sMined method (WGS 16S method) performs best amongst the 4 whole genome sequence based methods evaluated. Within other 3 methods theWebCARMA method slightly outperforms BlastBestHit and Megan, at all taxonomic levels except Phylum level.

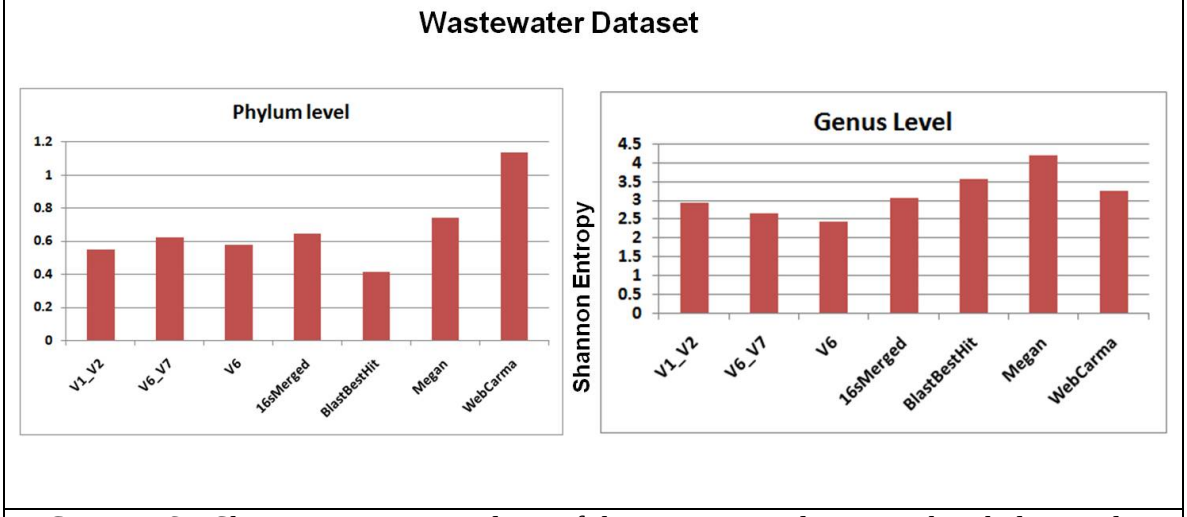


FIGURE 4.6a: Shannon Diversity indices of the wastewater dataset at the Phylum and Genus levels using the different taxonomic profiling methods is shown.

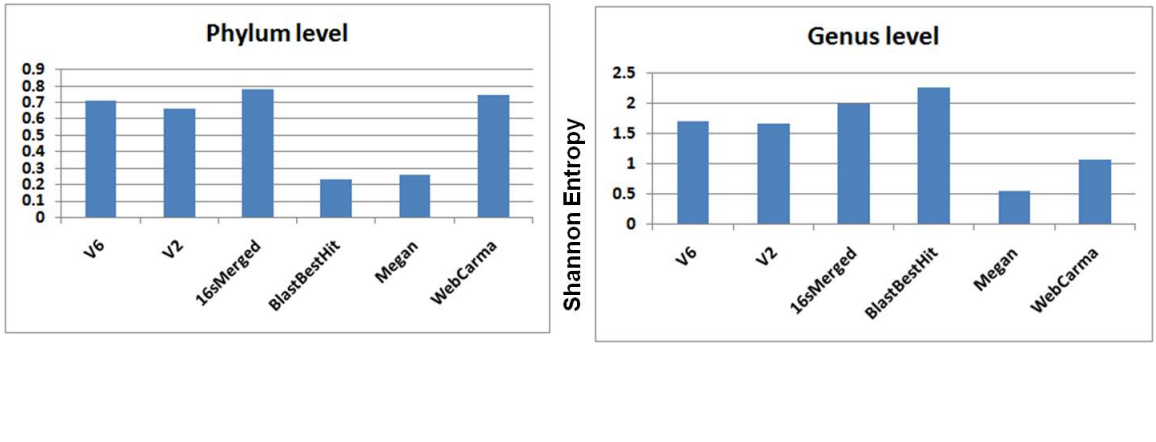


FIGURE 4.6b: Shannon Diversity indices of the human gut microbiome dataset at the Phylum and Genus levels using the different taxonomic profiling methods is shown.

CHAPTER 5: CONCLUSIONS

Microbes are ubiquitous and interest in microbes and their interactions with their habitats has been a focus of biological research from time immemorial. This interest has been renewed in recent times with rapid increase in sequencing technologies, resulting in decreasing sequencing costs, making the study of entire microbial communities easier and accessible to everyone. The consequence of this is a deluge of enormous amounts of data and novel and efficient bioinformatic tools are needed to decipher this complex data.

Availability of high-throughput technologies, tools and analytical methods helps us generate large quantities of data related to the microbial communities and help us understand these complex communities in a fraction of the time it took us decades ago. However, the biggest trade-off is that with increasing sample, sequence and data volumes, the experimental, technical and analysis pitfalls encountered are also amplified. Technical artifacts, unfortunately, pose a huge problem during metagenomic analyses. Following proper checks and balances throughout the course of such large scale studies, from the design to the data analysis stage is what is required to weed out the artifacts from the true biological effects. Since every step of metagenomic analysis involves choices (experimental, technological or analytical) to be made, and since the choice of method can have an effect on the results, there is a pressing need for standards to be established in this field.

During the course of this dissertation, we establish the value of using bioinformatics tools to understand complex ecosystems, not only by filtering out the unwanted artifacts and by helping make informed analysis choices, but also in reaching biologically important and accurate conclusions.

REFERENCES

1. Flores, G.E., et al., Sulfurihydrogenibium kristjanssonii sp. nov., a hydrogen- and sulfur-oxidizing thermophile isolated from a terrestrial Icelandic hot spring. Int J Syst Evol Microbiol, 2008. 58(Pt 5): p. 1153-8.

2. Brim, H., et al., Engineering Deinococcus radiodurans for metal remediation in radioactive mixed waste environments. Nat Biotechnol, 2000. 18(1): p. 85-90.

3. Edwards, K.J., W. Bach, and D.R. Rogers, Geomicrobiology of the ocean crust: a role for chemoautotrophic Fe-bacteria. Biol Bull, 2003. 204(2): p. 180-5.

4. Adesemoye, A.O. and J.W. Kloepper, Plant-microbes interactions in enhanced fertilizer-use efficiency. Appl Microbiol Biotechnol, 2009. 85(1): p. 1-12.

5. Hooper, L.V., T. Midtvedt, and J.I. Gordon, How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr, 2002. 22: p. 283-307.

6. Backhed, F., et al., Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U S A, 2007. 104(3): p. 979-84.

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

8. Backhed, F., et al., The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A, 2004. 101(44): p. 15718-23.

9. Backhed, F., et al., Host-bacterial mutualism in the human intestine. Science, 2005. 307(5717): p. 1915-20.

10. Venter, J.C., et al., Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science, 2004. 304(5667): p. 66.

11. Daniel, R., The metagenomics of soil. Nat Rev Micro, 2005. 3(6): p. 470-478.

12. Gill, S.R., et al., Metagenomic analysis of the human distal gut microbiome. Science, 2006. 312(5778): p. 1355-9.

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

14. Raes, J., K.U. Foerstner, and P. Bork, Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol, 2007. 10(5): p. 490-8.

15. Handelsman, J., et al., Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol, 1998. 5(10): p. R245-9.

16. Chen, K. and L. Pachter, Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. PLoS Computational Biology, 2005. 1(2).

17. Rodrigue, S., et al., Unlocking short read sequencing for metagenomics. PLoS One, 2010. 5(7): p. e11840.

18. Hugenholtz, P., B.M. Goebel, and N.R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol, 1998. 180(18): p. 4765-74.

19. Fleischmann, R.D., et al., Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 1995. 269(5223): p. 496-512.

20. Istrail, S., et al., Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci U S A, 2004. 101(7): p. 1916-21.

21. Venter, J.C., et al., The sequence of the human genome. Science, 2001. 291(5507): p. 1304-51.

22. Qin, J., et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010. 464(7285): p. 59-65.

23. Yooseph, S., et al., The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol, 2007. 5(3): p. e16.

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

25. Margulies, M., et al., Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005. 437(7057): p. 376-80.

26. Rohwer, F., Global phage diversity. Cell, 2003. 113(2): p. 141.

27. Hall, N., Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol, 2007. 210(Pt 9): p. 1518-25.

28. Tyson, G.W., et al., Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, 2003. 428: p. 37-43.

29. Salzberg, S.L. and J.A. Yorke, Beware of mis-assembled genomes. Bioinformatics, 2005. 21(24): p. 4320-1.

30. DeLong, E.F., et al., Community genomics among stratified microbial assemblages in the ocean's interior. Science, 2006. 311(5760): p. 496-503.

31. Edwards, R., et al., Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics, 2006. 7(1): p. 57.

32. Liu, Z., et al., Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res, 2007. 35(18): p. e120.

33. Woese, C.R. and G.E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A, 1977. 74(11): p. 5088-90.

34. Janda, J.M. and S.L. Abbott, 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol, 2007. 45(9): p. 2761-4.

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

36. Muyzer, G. and K. Smalla, Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie van Leeuwenhoek, 1998. 73(1): p. 127-141.

37. Stamper, D.M., M. Walch, and R.N. Jacobs, Bacterial Population Changes in a Membrane Bioreactor for Graywater Treatment Monitored by Denaturing Gradient Gel Electrophoretic Analysis of 16S rRNA Gene Fragments. Appl. Environ. Microbiol., 2003. 69(2): p. 852-860.

38. Wilson, K.H., et al., High-Density Microarray of Small-Subunit Ribosomal DNA Probes. Appl. Environ. Microbiol., 2002. 68(5): p. 2535-2541.

39. Andersson, A.F., et al., Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One, 2008. 3(7): p. e2836.

40. Blackall, L.L., et al., The use of 16S rDNA clone libraries to describe the microbial diversity of activated sludge communities. Water Science and Technology, 1998. 37(4-5): p. 451–454

41. Edwards, R.A., et al., Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics, 2006. 7: p. 57.

42. Tringe, S.G. and P. Hugenholtz, A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol, 2008. 11(5): p. 442-6.

43. Sogin, M.L., et al., Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A, 2006. 103(32): p. 12115-20.

44. Hamp, T.J., W.J. Jones, and A.A. Fodor, Effects of experimental choices and analysis noise on surveys of the "rare biosphere". Appl Environ Microbiol, 2009. 75(10): p. 3263-70.

45. Liu, W., et al., [Analysis of soil bacterial diversity by using the 16S rRNA gene library]. Wei Sheng Wu Xue Bao, 2008. 48(10): p. 1344-50.

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

47. Dethlefsen, L., et al., The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol, 2008. 6(11): p. e280.

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

49. Roesch, L., et al., Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME Journal, 2007. 1: p. 283-290.

50. Handelsman, J., Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev., 2004. 68(4): p. 669-685.

51. Baker, B.J., et al., Lineages of acidophilic archaea revealed by community genomic analysis. Science, 2006. 314(5807): p. 1933-5.

52. Turnbaugh, P.J., et al., A core gut microbiome in obese and lean twins. Nature, 2009. 457(7228): p. 480-4.

53. Sokol, H., et al., Specificities of the fecal microbiota in inflammatory bowel disease. Inflamm Bowel Dis, 2006. 12(2): p. 106-11.

54. Nishikawa, J., et al., Diversity of mucosa-associated microbiota in active and inactive ulcerative colitis. Scand J Gastroenterol, 2009. 44(2): p. 180-6.

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

56. Huycke, M.M. and H.R. Gaskins, Commensal bacteria, redox stress, and colorectal cancer: mechanisms and models. Exp Biol Med (Maywood), 2004. 229(7): p. 586-97.

57. Shen, X.J., Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. Gut Microbes, 2010. 1(3): p. 10.

58. Savage, D.C., Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol, 1977. 31: p. 107-33.

59. Ley, R.E., et al., Microbial ecology: human gut microbes associated with obesity. Nature, 2006. 444(7122): p. 1022-3.

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

61. Sundquist, A., et al., Bacterial flora-typing with targeted, chip-based Pyrosequencing. BMC Microbiol, 2007. 7: p. 108.

62. Keijser, B.J., et al., Pyrosequencing analysis of the oral microflora of healthy adults. J Dent Res, 2008. 87(11): p. 1016-20.

63. Grice, E.A., et al., A diversity profile of the human skin microbiota. Genome Res, 2008. 18(7): p. 1043-50.

64. Sears, C.L., A dynamic partnership: celebrating our gut flora. Anaerobe, 2005. 11(5): p. 247-51.

65. Guarner, F. and J.R. Malagelada, Gut flora in health and disease. Lancet, 2003. 361(9356): p. 512-9.

66. Dicksved, J., et al., Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J, 2008. 2(7): p. 716-27.

67. Dicksved, J., et al., Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls. J Med Microbiol, 2009. 58(Pt 4): p. 509-16.

68. Willing, B., et al., Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease. Inflamm Bowel Dis, 2009. 15(5): p. 653-60.

69. Turnbaugh, P.J. and J.I. Gordon, The core gut microbiome, energy balance and obesity. J Physiol, 2009. 587(Pt 17): p. 4153-8.

70. Turnbaugh, P.J., et al., An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 2006. 444(7122): p. 1027-31.

71. Cani, P.D. and N.M. Delzenne, Interplay between obesity and associated metabolic disorders: new insights into the gut microbiota. Curr Opin Pharmacol, 2009. 9(6): p. 737-43.

72. Tsukumo, D.M., et al., Translational research into gut microbiota: new horizons in obesity treatment. Arq Bras Endocrinol Metabol, 2009. 53(2): p. 139-44.

73. Moore, W.E. and L.H. Moore, Intestinal floras of populations that have a high risk of colon cancer. Appl Environ Microbiol, 1995. 61(9): p. 3202-7.

74. Powrie, F. and H. Uhlig, Animal models of intestinal inflammation: clues to the pathogenesis of inflammatory bowel disease. Novartis Found Symp, 2004. 263: p. 164-74; discussion 174-8, 211-8.

75. Rakoff-Nahoum, S. and R. Medzhitov, Role of the innate immune system and host-commensal mutualism. Curr Top Microbiol Immunol, 2006. 308: p. 1-18.

76. Rakoff-Nahoum, S., et al., Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell, 2004. 118(2): p. 229-41.

77. Andoh, A., et al., Recent advances in molecular approaches to gut microbiota in inflammatory bowel disease. Curr Pharm Des, 2009. 15(18): p. 2066-73.

78. Burcelin, R., et al., The gut microbiota ecology: a new opportunity for the treatment of metabolic diseases? Front Biosci, 2009. 14: p. 5107-17.

79. Ling, Z., et al., Analysis of Oral Microbiota in Children with Dental Caries by PCR-DGGE and Barcoded Pyrosequencing. Microb Ecol, 2010.

80. Fujimura, K.E., et al., Role of the gut microbiota in defining human health. Expert Rev Anti Infect Ther, 2010. 8(4): p. 435-54.

81. Mazmanian, S.K. and D.L. Kasper, The love-hate relationship between bacterial polysaccharides and the host immune system. Nat Rev Immunol, 2006. 6(11): p. 849-58.

82. Chow, J. and S.K. Mazmanian, A pathobiont of the microbiota balances host colonization and intestinal inflammation. Cell Host Microbe, 2010. 7(4): p. 265-76.

83. Boyle, P. and J. Ferlay, Cancer incidence and mortality in Europe, 2004. Ann Oncol, 2005. 16(3): p. 481-8.

84. Guilera, M., et al., Does physical activity modify the association between body mass index and colorectal adenomas? Nutr Cancer, 2005. 51(2): p. 140-5.

85. Bingham, S.A., Diet and colorectal cancer prevention. Biochem Soc Trans, 2000. 28(2): p. 12-6.

86. Keku, T.O., et al., Rectal mucosal proliferation, dietary factors, and the risk of colorectal adenomas. Cancer Epidemiol Biomarkers Prev, 1998. 7(11): p. 993-9.

87. Green, G.L., et al., Molecular characterization of the bacteria adherent to human colorectal mucosa. J Appl Microbiol, 2006. 100(3): p. 460-9.

88. Scanlan, P.D., et al., Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. Environ Microbiol, 2008. 10(3): p. 789-98.

89. Guldberg, P., et al., Single-step DGGE-based mutation scanning of the p53 gene: application to genetic diagnosis of colorectal cancer. Hum Mutat, 1997. 9(4): p. 348-55.

90. Keku, T.O., et al., Insulin resistance, apoptosis, and colorectal adenoma risk. Cancer Epidemiol Biomarkers Prev, 2005. 14(9): p. 2076-81.

91. Leek, J.T., et al., Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet, 2010. 11(10): p. 733-9.

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

93. Li, S. and H.H. Chou, LUCY2: an interactive DNA sequence quality trimming and vector removal tool. Bioinformatics, 2004. 20(16): p. 2865-6.

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

95. Benjamini, Y. and Y. Hochberg, A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society., 1995. Series B (Methodological) Vol. 57(No. 1): p. 12.

96. Lozupone, C. and R. Knight, UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol, 2005. 71(12): p. 8228-35.

97. Livak, K.J. and T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 2001. 25(4): p. 402-8.

98. Zhang, H., et al., Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A, 2009. 106(7): p. 2365-70.

99. Chichlowski, M. and L.P. Hale, Bacterial-mucosal interactions in inflammatory bowel disease: an alliance gone bad. Am J Physiol Gastrointest Liver Physiol, 2008. 295(6): p. G1139-49.

100. Jones, M., et al., Helicobacter pylori in colorectal neoplasms: is there an aetiological relationship? World J Surg Oncol, 2007. 5: p. 51.

101. Burnett-Hartman, A.N., P.A. Newcomb, and J.D. Potter, Infectious agents and colorectal cancer: a review of Helicobacter pylori, Streptococcus bovis, JC virus, and human papillomavirus. Cancer Epidemiol Biomarkers Prev, 2008. 17(11): p. 2970-9.

102. Zumkeller, N., et al., Helicobacter pylori infection and colorectal cancer risk: a meta-analysis. Helicobacter, 2006. 11(2): p. 75-80.

103. Abbolito, M.R., et al., The association of Helicobacter pylori infection with low levels of urea and pH in the gastric juices. Ital J Gastroenterol, 1992. 24(7): p. 389-92.

104. Chen, G., et al., Helicobacter pylori survival in gastric mucosa by generation of a pH gradient. Biophys J, 1997. 73(2): p. 1081-8.

105. Tanaka, N., et al., Flagellin from an incompatible strain of Acidovorax avenae mediates H2O2 generation accompanying hypersensitive cell death and expression of PAL, Cht-1, and PBZ1, but not of Lox in rice. Mol Plant Microbe Interact, 2003. 16(5): p. 422-8.

106. Takakura, Y., et al., Expression of a bacterial flagellin gene triggers plant immune responses and confers disease resistance in transgenic rice plants. Mol Plant Pathol, 2008. 9(4): p. 525-9.

107. Gibson, G.R. and M.B. Roberfroid, Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. J Nutr, 1995. 125(6): p. 1401-12.

108. Macfarlane, S., G.T. Macfarlane, and J.H. Cummings, Review article: prebiotics in the gastrointestinal tract. Aliment Pharmacol Ther, 2006. 24(5): p. 701-14.

109. Duncan, S.H., P. Louis, and H.J. Flint, Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Appl Environ Microbiol, 2004. 70(10): p. 5810-7.

110. Keku, T.O., et al., Local IGFBP-3 mRNA expression, apoptosis and risk of colorectal adenomas. BMC Cancer, 2008. 8: p. 143.

111. Spielman, R.S., et al., Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet, 2007. 39(2): p. 226-31.

112. Petricoin, E.F., et al., Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 2002. 359(9306): p. 572-7.

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

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

115. Mahowald, M.A., et al., Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc Natl Acad Sci U S A, 2009. 106(14): p. 5859-64.

116. Spencer, M.D., et al., Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. Gastroenterology, 2011. 140(3): p. 976-86.

117. Sanapareddy, N., et al., Molecular diversity of a North Carolina wastewater treatment plant as revealed by pyrosequencing. Appl Environ Microbiol, 2009. 75(6): p. 1688-96.

118. Barker, A.V. and G.M. Bryson, Bioremediation of heavy metals and organic toxicants by composting. ScientificWorldJournal, 2002. 2: p. 407-20.

119. Cohen, Y., Bioremediation of oil by marine microbial mats. Int Microbiol, 2002. 5(4): p. 189-93.

120. Fernandez-Alvarez, P., et al., Evaluation of biodiesel as bioremediation agent for the treatment of the shore affected by the heavy oil spill of the Prestige. J Hazard Mater, 2007. 147(3): p. 914-22.

121. Bitton, G., Wastewater Microbiology. 1999, New York: Wiley-Liss.

122. Howarth, R., et al., Phylogenetic relationships of filamentous sulfur bacteria (Thiothrix spp. and Eikelboom type 021N bacteria) isolated from wastewater-treatment plants and description of Thiothrix eikelboomii sp. nov., Thiothrix unzii sp. nov., Thiothrix fructosivorans sp. nov. and Thiothrix defluvii sp. nov. Int J Syst Bacteriol, 1999. 49 Pt 4: p. 1817-27.

123. Jones, W., P. Wilderer, and E. Schroeder, Operation of a Three-Stage SBR System for Nitrogen Removal from Wastewater Research Journal of the Water Pollution Control Federation, 1990. 62(3): p. 268-274.

124. Garcia Martin, H., et al., Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol, 2006. 24(10): p. 1263-9.

125. Amann, R., H. Lemmer, and M. Wagner, Monitoring the community structure of wastewater treatment plants: a comparison of old and new techniques. FEMS Microbiology Ecology, 1998. 25(3): p. 205-215.

126. Wagner, M., et al., Microbial community composition and function in wastewater treatment plants. Antonie van Leeuwenhoek, 2002. 81(1): p. 665-680.

127. Boon, N., et al., Evaluation of nested PCR-DGGE(denaturing gradient gel electrophoresis) with group-specific 16 S rRNA primers for the analysis of

bacterial communities from different wastewater treatment plants. FEMS Microbiology Ecology, 2002. 39(2): p. 101-112.

128. Layton, A.C., et al., Quantification of Hyphomicrobium Populations in Activated Sludge from an Industrial Wastewater Treatment System as Determined by 16S rRNA Analysis. Appl. Environ. Microbiol., 2000. 66(3): p. 1167-1174.

129. Gilbride, K.A. and R.R. Fulthorpe, A survey of the composition and diversity of bacterial populations in bleached kraft pulp-mill wastewater secondary treatment systems. Canadian Journal of Microbiology, 2004. 50(8): p. 633–644.

130. Coskuner, G. and T.P. Curtis, In situ characterization of nitrifiers in an activated sludge plant: detection of Nitrobacter Spp. Journal of Applied Microbiology, 2002. 93(3): p. 431-437.

131. Seviour, R.J., T. Mino, and M. Onuki, The microbiology of biological phosphorus removal in activated sludge systems. FEMS Microbiology Ecology, 2003. 27(1): p. 99-127.

132. Jeon, C.O., D.S. Lee, and J.M. Park, Microbial communities in activated sludge performing enhanced biological phosphorus removal in a sequencing batch reactor. Water Research, 2003. 37(9): p. 2195-2205.

133. Bond, P.L., et al., Identification of Some of the Major Groups of Bacteria in Efficient and Nonefficient Biological Phosphorus Removal Activated Sludge Systems. Appl. Environ. Microbiol., 1999. 65(9): p. 4077-4084.

134. Juretschko, S., et al., Combined Molecular and Conventional Analyses of Nitrifying Bacterium Diversity in Activated Sludge: Nitrosococcus mobilis and Nitrospira-Like Bacteria as Dominant Populations. Appl. Environ. Microbiol., 1998. 64(8): p. 3042-3051.

135. Otawa, K., et al., Molecular analysis of ammonia-oxidizing bacteria community in intermittent aeration sequencing batch reactors used for animal wastewater treatment. Environmental Microbiology, 2006. 8(11): p. 1985-1996.

136. Gilbert, Y., Y.L. Bihan, and P. Lessard, Acetylene blockage technique as a tool to determine denitrification potential of a biomass fixed on an organic media treating wastewater. J. Environ. Eng. Sci, 2006. 5(5): p. 437-442.

137. Beline, F., et al., Application of the 15N technique to determine the contributions of nitrification and denitrification to the flux of nitrous oxide from aerated pig slurry. Water Res., 2001. 65(11): p. 2774-2778.

138. Zheng, D., Quantification of Methanosaeta Species in Anaerobic Bioreactors Using Genus-and Species-Specific Hybridization Probes. Microbial Ecology, 2000. 39(3): p. 246-262.

139. Jupraputtasri, W., et al., Use of an alternative Archaea-specific probe for methanogen detection. Journal of Microbiological Methods, 2005. 61(1): p. 95-104.

140. Oerther, D.B., et al., Quantifying filamentous microorganisms in activated sludge before, during, and after an incident of foaming by oligonucleotide probe hybridizations and antibody staining. Water Res, 2001. 35(14): p. 3325-36.

141. de los Reyes, F.L., D. Rothauszky, and L. Raskin, Microbial Community Structures in Foaming and Nonfoaming Full-Scale Wastewater Treatment Plants. Water Environment Research, 2002. 74(5): p. 437-449.

142. Rusch, D.B., et al., The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol, 2007. 5(3): p. e77.

143. Mavromatis, K., et al., Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods, 2007. 4(6): p. 495-500.

144. Besemer, J. and M. Borodovsky, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res, 2005. 33(Web Server issue): p. W451-4.

145. Cole, J.R., et al., The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucl. Acids Res., 2007. 35(Database issue): p. D169-D172.

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

147. Hamp, T.J., W.J. Jones, and A.A. Fodor, Targeting the V1-V2 region of the 16S rRNA gene yields improved measures of microbial community composition in a pyrosequencing survey of a wastewater treatment plant. submitted.

148. Huber, J.A., et al., Microbial population structures in the deep marine biosphere. Science, 2007. 318(5847): p. 97-100.

149. Pauli, W., K. Jax, and S. Berger, Protozoa in Wastewater Treatment: Function and Importance, in Biodegradation and Persistance. 2001, Springer Berlin / Heidelberg. p. 203-252.

150. Tringe, S.G., et al., Comparative Metagenomics of Microbial Communities. Science, 2005. 308(5721): p. 554-557.

151. Dinsdale, E.A., et al., Functional metagenomic profiling of nine biomes. Nature, 2008. 452(7187): p. 629-32.

152. Overbeek, R., et al., The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res, 2005. 33(17): p. 5691-702.

153. Aziz, R.K., et al., The RAST Server: rapid annotations using subsystems technology. BMC Genomics, 2008. 9: p. 75.

154. Huson, D.H., et al., MEGAN analysis of metagenomic data. Genome Res, 2007. 17(3): p. 377-86.

155. Krause, L., et al., Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res, 2008. 36(7): p. 2230-9.

156. Janda, I. and K. Mikulik, Preparation and sequencing of the cloacin fragment of Streptomyces aureofaciens 16S RNA. Biochem Biophys Res Commun, 1986. 137(1): p. 80-6.

157. Weisburg, W.G., et al., 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol, 1991. 173(2): p. 697-703.

158. Case, R.J., et al., Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol, 2007. 73(1): p. 278-88.

159. Baker, G.C., J.J. Smith, and D.A. Cowan, Review and re-analysis of domain-specific 16S primers. J Microbiol Methods, 2003. 55(3): p. 541-55.

160. Achtman, M. and M. Wagner, Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol, 2008. 6(6): p. 431-40.

161. Gerlach, W., et al., WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics, 2009. 10: p. 430.

162. Schreiber, F., et al., Treephyler: fast taxonomic profiling of metagenomes. Bioinformatics, 2010. 26(7): p. 960-1.

163. McHardy, A.C., et al., Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods, 2007. 4(1): p. 63-72.

164. Diaz, N.N., et al., TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics, 2009. 10: p. 56.

165. Brady, A. and S.L. Salzberg, Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods, 2009. 6(9): p. 673-676.

166. DeSantis, T.Z., Jr., et al., NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucl. Acids Res., 2006. 34(suppl_2): p. W394-399.

167. Manichanh, C., et al., A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. Nucleic Acids Res, 2008. 36(16): p. 5180-8.

168. Monzoorul Haque, M., et al., SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics, 2009. 25(14): p. 1722-30.

APPENDIX A: SUPPLEMENTARY MATERIALS FOR CHAPTER 1

Supplementary Tables

Supplementary Table 1: Wilcoxon-tests on log-normalized abundances of all phyla in cases (33 subjects) vs. controls (38 subjects). Only phyla which have at least 1 sequence assigned to them in 25% of the samples are shown. The direction of change shows the relative abundance in cases compared to controls. Wilcoxon p-Values were corrected for multiple testing[1] using (n*p)/R where n= total number of taxa tested, p= raw p-Value and R= sorted Rank of the taxon. *While the sequences classified to Cyanobacteria may in fact originate from plastids or from non-Cyanobacteria, other human and animal gut studies[2] have also observed sequences classified to Cyanobacteria.

Phylum Name	Wilcoxon p-Value	Rank	(n*p)/R	Direction
TM7	0.00020	1	0.00180	Up
Cyanobacteria*	0.00220	2	0.00990	Up
Verrucomicrobia	0.00610	3	0.01830	Up
Firmicutes	0.04740	4	0.10665	Down
Acidobacteria	0.06010	5	0.10818	Up
Fusobacteria	0.17740	6	0.26610	Up
Proteobacteria	0.18110	7	0.23284	Up
Actinobacteria	0.31030	8	0.34909	Up
Bacteroidetes	0.83560	9	0.83560	Up

Supplementary Table 2: Wilcoxon-tests on log-normalized abundances of genera in cases (33 subjects) vs. controls (38 subjects). Only genera which have at least 1 sequence assigned to them in 25% of the samples are shown. The direction of change shows the relative abundance in cases compared to controls. Wilcoxon p-Values were corrected for multiple testing[1] using (n*p)/R where n= total number of taxa tested, p= raw p-Value and R= sorted Rank of the taxon.

Genus	Wilcoxon p-Value	Rank	(n*p)/R	Direction
Helicobacter	0.00003	1	0.00290	Up
Aquabacterium	0.00005	2	0.00270	Up
Weissella	0.00026	3	0.00870	Up
Lactococcus	0.00070	4	0.01748	Up
Acidovorax	0.00083	5	0.01666	Up
Turicibacter	0.00128	6	0.02138	Up
Lactobacillus	0.00134	7	0.01917	Up
Sphingobium	0.00137	8	0.01715	Up
Cloacibacterium	0.00145	9	0.01611	Up
Stenotrophomonas	0.00171	10	0.01709	Up
Succinivibrio	0.00261	11	0.02374	Up
Azonexus	0.00324	12	0.02702	Up
Leuconostoc	0.00326	13	0.02504	Up
Delftia	0.00385	14	0.02752	Up
Dechloromonas	0.00401	15	0.02673	Up
Akkermansia	0.00595	16	0.03717	Up
Bryantella	0.00682	17	0.04012	Up
Acinetobacter	0.00711	18	0.03947	Up
Agrobacterium	0.00882	19	0.04643	Up
Streptococcus	0.01006	20	0.05028	Down
Bacillaceae_1	0.01384	21	0.06590	Up
Allobaculum	0.01408	22	0.06400	Up
Serratia	0.01620	23	0.07044	Up
Rubrobacterineae	0.01729	24	0.07206	Up
Chryseobacterium	0.01947	25	0.07788	Up
Micrococcineae	0.01948	26	0.07493	Up
Pantoea	0.02126	27	0.07873	Up

Gp2	0.02315	28	0.08267	Up
Pseudomonas	0.02367	29	0.08161	Up
Exiguobacterium	0.02493	30	0.08310	Up
Gp1	0.02806	31	0.09051	Up
Pseudoxanthomonas	0.04403	32	0.13759	Up
Dorea	0.04758	33	0.14418	Down
Novosphingobium	0.04910	34	0.14441	Up
Sutterella	0.05041	35	0.14403	Up
Bifidobacteriaceae	0.05077	36	0.14102	Down
Chryseomonas	0.05792	37	0.15654	Up
Comamonas	0.07497	38	0.19730	Up
Carnobacteriaceae_1	0.07831	39	0.20080	Up
Alistipes	0.08070	40	0.20175	Up
Bacteroides	0.09360	41	0.22829	Down
Staphylococcus	0.10208	42	0.24304	Up
Variovorax	0.10572	43	0.24585	Up
Flavimonas	0.11058	44	0.25131	Up
Shinella	0.12952	45	0.28783	Up
Syntrophococcus	0.13651	46	0.29676	Up
Methylobacterium	0.13766	47	0.29290	Up
Roseburia	0.15451	48	0.32189	Up
Enterobacter	0.15715	49	0.32072	Up
Erwinia	0.16696	50	0.33392	Up
Rheinheimera	0.17078	51	0.33486	Down
Prevotella	0.19727	52	0.37936	Up
Succinispira	0.20400	53	0.38491	Up
Pedobacter	0.23060	54	0.42704	Up
Fusobacterium	0.23880	55	0.43419	Up
Sphingomonas	0.25308	56	0.45192	Up
Bradyrhizobium	0.25361	57	0.44492	Down
Propionibacterineae	0.26446	58	0.45596	Up
Burkholderia	0.26620	59	0.45119	Up
Veillonella	0.28595	60	0.47659	Down
Vibrio	0.28683	61	0.47022	Down
Papillibacter	0.28810	62	0.46468	Up

Marinomonas	0.31275	63	0.49643	Down
Bilophila	0.40399	64	0.63123	Up
Gemella	0.40841	65	0.62832	Up
Enhydrobacter	0.44562	66	0.67518	Up
Anaerococcus	0.45866	67	0.68456	Up
Pseudoalteromonas	0.47369	68	0.69660	Down
Finegoldia	0.49275	69	0.71413	Down
Haemophilus	0.49499	70	0.70712	Down
Butyrivibrio	0.52466	71	0.73896	Up
Coprococcus	0.53663	72	0.74532	Up
Clostridiaceae_1	0.57343	73	0.78553	Up
Ruminococcaceae_Incertae_Sedis	0.59101	74	0.79867	Up
Paracoccus	0.61333	75	0.81777	Up
Anaerotruncus	0.64579	76	0.84973	Down
Parabacteroides	0.64883	77	0.84264	Up
Lachnospiraceae_Incertae_Sedis	0.68417	78	0.87714	Up
Citrobacter	0.68862	79	0.87167	Up
Coprobacillus	0.69082	80	0.86352	Down
Desulfovibrio	0.71148	81	0.87837	Down
Shigella	0.72933	82	0.88943	Down
Actinomycineae	0.74703	83	0.90004	Down
Uruburuella	0.75252	84	0.89586	Down
Corynebacterineae	0.78329	85	0.92152	Down
Megamonas	0.84097	86	0.97787	Down
Aeromonas	0.85775	87	0.98592	Down
Holdemania	0.86825	88	0.98665	Up
Subdoligranulum	0.87174	89	0.97948	Up
Coriobacterineae	0.87710	90	0.97456	Down
Ralstonia	0.88637	91	0.97403	Up
Erysipelotrichaceae_Incertae_Sedis	0.89520	92	0.97304	Up
Allomonas	0.91827	93	0.98739	Down
Peptostreptococcaceae_Incertae_Sedis	0.93100	94	0.99043	Up
Brevundimonas	0.94692	95	0.99676	Down
Carnobacteriaceae_2	0.94786	96	0.98736	Up
Anaerovorax	0.96308	97	0.99286	Down

Faecalibacterium	0.97701	98	0.99695	Up
Ruminococcus	0.98616	99	0.99612	Up
Dialister	0.99025	100	0.99025	Up

Supplementary Table 3: Wilcoxon-tests on log-normalized abundances of OTUs (97%) in cases (33 subjects) vs. controls (38 subjects). Only OTUs which have at least 1 sequence assigned to them in 25% of the samples are shown. RDP classification of consensus sequences at genus level shown. Wilcoxon p-Values were corrected for multiple testing[1] using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted Rank of the taxon.

OtuName	Wilcoxon p-Value	Rank	n*p/R	Direction	RDP genus level Assignment
OTU72	0.000084	1	0.031257	Up	Aquabacterium
OTU226	0.000085	2	0.015686	Up	Rikenella
OTU200	0.000087	3	0.010705	Up	Helicobacter
OTU432	0.000111	4	0.010297	Up	Paludibacter
OTU285	0.000137	5	0.010167	Up	Butyrivibrio
OTU157	0.000139	6	0.008578	Up	Marinilabilia
OTU240	0.000318	7	0.016856	Up	Weissella
OTU370	0.000384	8	0.017786	Up	Lactobacillus
OTU284	0.000424	9	0.017486	Down	Rubritepida
OTU22	0.00043	10	0.015937	Up	Acidovorax
OTU96	0.000484	11	0.016326	Up	Diaphorobacter
OTU119	0.000579	12	0.017915	Up	Lachnobacterium
OTU213	0.000679	13	0.019378	Up	Lactococcus
OTU73	0.000703	14	0.018642	Up	Lactococcus
OTU306	0.000821	15	0.020303	Down	Oligotropha
OTU373	0.000896	16	0.020772	Up	Sporobacter
OTU501	0.000947	17	0.020667	Up	Ruminococcaceae Incertae Sedis
OTU37	0.001006	18	0.020743	Up	Cloacibacterium
OTU109	0.001008	19	0.019674	Up	Turicibacter
OTU100	0.001258	20	0.023329	Up	Xylanibacter
OTU122	0.001335	21	0.023579	Up	Prevotella
OTU46	0.001398	22	0.023569	Up	Bacillaceae 1
OTU525	0.001497	23	0.024146	Up	Catonella
OTU70	0.001582	24	0.02446	Up	Sphingobium
OTU91	0.001641	25	0.024351	Up	Lactobacillus
OTU75	0.001703	26	0.024306	Up	Stenotrophomonas
OTU328	0.00179	27	0.02459	Up	Parasporobacterium

OTU309	0.002063	28	0.027333	Up	Paludibacter
OTU230	0.002084	29	0.026658	Up	Butyrivibrio
OTU371	0.002129	30	0.02633	Up	Comamonas
OTU177	0.002213	31	0.026484	Up	Butyrivibrio
OTU136	0.002304	32	0.026712	Up	Micrococcineae
OTU357	0.002384	33	0.026803	Up	Coprococcus
OTU387	0.002449	34	0.026723	Up	Coprococcus
OTU124	0.002547	35	0.026996	Up	Lactobacillus
OTU38	0.002829	36	0.029152	Up	Pseudomonas
OTU56	0.002884	37	0.028914	Up	Delftia
OTU202	0.002913	38	0.028437	Up	Lachnospiraceae Incertae Sedis
OTU133	0.002963	39	0.028182	Up	Faecalibacterium
OTU242	0.003059	40	0.028371	Up	Coriobacterineae
OTU189	0.00349	41	0.031576	Up	Acidovorax
OTU439	0.003755	42	0.033171	Down	Algibacter
OTU265	0.003802	43	0.032805	Up	Sphingomonas
OTU139	0.003893	44	0.032827	Up	Azonexus
OTU95	0.004005	45	0.03302	Up	Ruminococcus
OTU23	0.004051	46	0.032674	Up	Lachnospiraceae Incertae Sedis
OTU59	0.004084	47	0.032241	Up	Acinetobacter
OTU502	0.004279	48	0.033077	Up	Paludibacter
OTU64	0.004323	49	0.032735	Up	Erwinia
OTU454	0.004669	50	0.034641	Up	Paludibacter
OTU286	0.005422	51	0.039446	Up	Hallella
OTU464	0.005427	52	0.038721	Up	Marinilabilia
OTU161	0.006285	53	0.043997	Up	Prevotella
OTU423	0.007065	54	0.048543	Up	Parasporobacterium
OTU53	0.007612	55	0.051345	Up	Succinivibrio
OTU239	0.007843	56	0.051957	Up	Succinispira
OTU319	0.008701	57	0.056633	Up	Agrobacterium
OTU193	0.008755	58	0.056004	Up	Xylanibacter
OTU61	0.009098	59	0.057207	Up	Papillibacter
OTU365	0.009827	60	0.060762	Up	Succinispira
OTU437	0.010114	61	0.061514	Up	Marinilabilia
OTU225	0.010608	62	0.063477	Up	Prevotella
OTU366	0.01081	63	0.063657	Up	Coprococcus
OTU92	0.01095	64	0.063478	Up	Rubrobacterineae
OTU463	0.01103	65	0.062958	Up	Lachnospiraceae Incertae Sedis
OTU97	0.011294	66	0.063484	Up	Pseudomonas

OTU21	0.011865	67	0.065699	Up	Finegoldia
OTU149	0.012682	68	0.069192	Down	Haemophilus
OTU241	0.013048	69	0.070156	Up	Chryseobacterium
OTU250	0.013254	70	0.070246	Up	Paludibacter
OTU210	0.013651	71	0.071332	Up	Allobaculum
OTU347	0.013893	72	0.071586	Down	Vitellibacter
OTU191	0.014678	73	0.074597	Up	Subdoligranulum
OTU404	0.014845	74	0.074425	Up	Hallella
OTU396	0.014935	75	0.073878	Up	Coprococcus
OTU345	0.01502	76	0.073319	Up	Butyrivibrio
OTU401	0.015426	77	0.074324	Up	Alistipes
OTU67	0.015821	78	0.075251	Up	Lactobacillus
OTU407	0.016533	79	0.077644	Up	Turicibacter
OTU313	0.016785	80	0.077842	Up	Enterobacter
OTU353	0.017139	81	0.0785	Up	Dorea
OTU418	0.019841	82	0.08977	Up	Stenotrophomonas
OTU393	0.020465	83	0.091478	Up	Micrococcineae
OTU120	0.020843	84	0.092056	Up	Micrococcineae
OTU413	0.021269	85	0.092833	Up	Subdoligranulum
OTU341	0.021427	86	0.092433	Up	Prevotella
OTU93	0.021869	87	0.093258	Up	Alistipes
OTU186	0.022338	88	0.094173	Up	Faecalibacterium
OTU79	0.022545	89	0.093981	Up	Lachnospiraceae Incertae Sedis
OTU197	0.023847	90	0.098304	Up	Lactobacillus
OTU219	0.024265	91	0.098928	Up	Rikenella
OTU86	0.02429	92	0.097951	Up	Fusobacterium
OTU297	0.0273	93	0.108905	Up	Bacillaceae 1
OTU442	0.02802	94	0.110588	Up	Roseburia
OTU389	0.028617	95	0.111759	Up	Parabacteroides
OTU352	0.028801	96	0.111304	Down	Saprospira
OTU49	0.031048	97	0.118749	Up	Sutterella
OTU329	0.032674	98	0.123693	Down	Methanohalobium
OTU176	0.033016	99	0.123727	Up	Erwinia
OTU484	0.033734	100	0.125152	Down	Effluviibacter
OTU569	0.033751	101	0.123975	Up	Erwinia
OTU66	0.034683	102	0.126152	Down	Streptococcus
OTU391	0.03501	103	0.126103	Up	Aquiflexum
OTU356	0.036933	104	0.131753	Up	Novosphingobium
OTU11	0.041357	105	0.146129	Up	Bacteroides
OTU330	0.04391	106	0.153686	Up	Coriobacterineae

OTU361	0.04391	107	0.152249	Up	Succinivibrio
OTU113	0.044104	108	0.151507	Up	Rikenella
OTU45	0.04423	109	0.150544	Down	Xenohaliotis
OTU471	0.045642	110	0.153937	Up	Lachnospiraceae Incertae Sedis
OTU247	0.047313	111	0.158135	Up	Xylanibacter
OTU283	0.050651	112	0.16778	Up	Anaerophaga
OTU128	0.055374	113	0.181802	Up	Prevotella
OTU270	0.056309	114	0.183252	Up	Succinispira
OTU57	0.061822	115	0.199442	Down	Lachnospiraceae Incertae Sedis
OTU77	0.06775	116	0.216684	Up	Coprococcus
OTU138	0.068101	117	0.215945	Down	Simkania
OTU491	0.068451	118	0.215214	Up	Clostridiaceae 1
OTU169	0.069264	119	0.215941	Down	Streptococcus
OTU207	0.070648	120	0.218419	Up	Succinispira
OTU237	0.072858	121	0.223392	Up	Prevotella
OTU499	0.075097	122	0.22837	Down	Lachnospiraceae Incertae Sedis
OTU14	0.07526	123	0.227004	Up	Erysipelotrichaceae Incertae Sedis
OTU417	0.07743	124	0.231665	Up	Lachnobacterium
OTU111	0.080236	125	0.23814	Up	Peptostreptococcaceae Incertae Sedis
OTU322	0.080575	126	0.237249	Up	Roseburia
OTU244	0.081081	127	0.236857	Up	Prevotella
OTU350	0.083008	128	0.240595	Up	Coprococcus
OTU159	0.084952	129	0.244319	Up	Faecalibacterium
OTU224	0.088054	130	0.251292	Up	Prevotella
OTU338	0.09269	131	0.262503	Up	Micrococcineae
OTU376	0.093281	132	0.262177	Up	Methylobacterium
OTU254	0.093506	133	0.260833	Down	Lachnospiraceae Incertae Sedis
OTU36	0.094305	134	0.261099	Up	Bacteroides
OTU8	0.095901	135	0.263551	Down	Dorea
OTU326	0.096151	136	0.262295	Down	Lachnospiraceae Incertae Sedis
OTU282	0.104442	137	0.282832	Down	Streptococcus
OTU264	0.107146	138	0.288052	Up	Comamonas
OTU26	0.11087	139	0.29592	Down	Dorea
OTU137	0.1132	140	0.299979	Up	Prevotella
OTU222	0.116058	141	0.305373	Up	Prevotella
OTU85	0.117436	142	0.306821	Up	Bacteroides
OTU397	0.12782	143	0.331617	Up	Peptostreptococcaceae Incertae Sedis

OTU167	0.129522	144	0.333699	Up	Allobaculum
OTU420	0.13338	145	0.341269	Up	Dorea
OTU474	0.13338	146	0.338931	Up	Sphingobium
OTU29	0.137289	147	0.346491	Down	Lachnospiraceae Incertae Sedis
OTU144	0.138737	148	0.347779	Down	Dorea
OTU172	0.140932	149	0.350912	Down	Marinilabilia
OTU409	0.141562	150	0.350129	Up	Alkalilimnicola
OTU68	0.145429	151	0.357313	Up	Dorea
OTU216	0.146992	152	0.358776	Up	Sphingomonas
OTU421	0.150949	153	0.366028	Down	Streptococcus
OTU476	0.157687	154	0.379882	Down	Streptococcus
OTU519	0.159874	155	0.382665	Up	Catonella
OTU143	0.160715	156	0.382213	Down	Lachnospiraceae Incertae Sedis
OTU275	0.160841	157	0.380078	Up	Lachnospiraceae Incertae Sedis
OTU206	0.161316	158	0.378785	Up	Paludibacter
OTU419	0.161556	159	0.376965	Up	Micrococcineae
OTU1	0.163025	160	0.378015	Down	Bacteroides
OTU248	0.16912	161	0.389711	Up	Lachnospiraceae Incertae Sedis
OTU134	0.169695	162	0.388622	Down	Ruminococcaceae Incertae Sedis
OTU141	0.174538	163	0.397262	Up	Faecalibacterium
OTU368	0.176676	164	0.399676	Up	Ruminococcaceae Incertae Sedis
OTU205	0.17885	165	0.402142	Up	Erysipelotrichaceae Incertae Sedis
OTU300	0.17925	166	0.400614	Down	Lachnospiraceae Incertae Sedis
OTU152	0.183253	167	0.407108	Down	Faecalibacterium
OTU82	0.189641	168	0.418791	Up	Roseburia
OTU28	0.194628	169	0.427261	Down	Bacteroides
OTU299	0.195265	170	0.426137	Up	Lachnospiraceae Incertae Sedis
OTU135	0.19551	171	0.424178	Up	Clostridiaceae 1
OTU267	0.197149	172	0.425246	Up	Parabacteroides
OTU249	0.197702	173	0.423974	Up	Faecalibacterium
OTU334	0.205736	174	0.438667	Up	Citrobacter
OTU34	0.206355	175	0.437473	Down	Dorea
OTU192	0.212037	176	0.446964	Up	Sphingomonas
OTU153	0.213057	177	0.446576	Up	Roseburia
OTU266	0.214087	178	0.446215	Down	Bacteroides
OTU87	0.215609	179	0.446876	Up	Propionibacterineae

OTU235	0.224633	180	0.462994	Up	Desulfovibrio
OTU50	0.226155	181	0.463556	Up	Sutterella
OTU33	0.229786	182	0.468411	Down	Lachnospiraceae Incertae Sedis
OTU90	0.231703	183	0.469737	Up	Lachnospiraceae Incertae Sedis
OTU204	0.231703	184	0.467184	Up	Dialister
OTU395	0.236361	185	0.474	Up	Subdoligranulum
OTU317	0.237329	186	0.473383	Up	Prevotella
OTU203	0.238017	187	0.472215	Down	Rheinheimera
OTU165	0.23893	188	0.471505	Up	Alistipes
OTU303	0.245272	189	0.481459	Down	Faecalibacterium
OTU15	0.246531	190	0.481385	Up	Roseburia
OTU127	0.246632	191	0.479061	Down	Lachnospiraceae Incertae Sedis
OTU412	0.248001	192	0.47921	Up	Sphingomonas
OTU178	0.250803	193	0.482114	Up	Lachnospiraceae Incertae Sedis
OTU195	0.252465	194	0.482808	Down	Pseudoalteromonas
OTU162	0.255823	195	0.486719	Down	Veillonella
OTU154	0.260826	196	0.493707	Down	Faecalibacterium
OTU190	0.260891	197	0.491324	Up	Ruminococcaceae Incertae Sedis
OTU74	0.263322	198	0.493397	Up	Ruminococcus
OTU425	0.264265	199	0.492674	Up	Enhydrobacter
OTU118	0.26768	200	0.496547	Up	Burkholderia
OTU83	0.268729	201	0.496012	Down	Dorea
OTU188	0.269309	202	0.494622	Down	Lachnospiraceae Incertae Sedis
OTU156	0.275877	203	0.504188	Up	Lachnospiraceae Incertae Sedis
OTU146	0.277131	204	0.503998	Down	Vibrio
OTU84	0.277838	205	0.50282	Down	Marinomonas
OTU3	0.286165	206	0.515375	Down	Lachnospiraceae Incertae Sedis
OTU170	0.2869	207	0.514203	Down	Bacteroides
OTU5	0.293459	208	0.52343	Up	Sphingomonas
OTU19	0.296777	209	0.526814	Up	Syntrophococcus
OTU142	0.301855	210	0.533278	Down	Lachnospiraceae Incertae Sedis
OTU307	0.303841	211	0.534242	Up	Megamonas
OTU360	0.310287	212	0.543003	Down	Faecalibacterium
OTU227	0.314679	213	0.548103	Down	Lachnospiraceae Incertae Sedis
OTU145	0.31593	214	0.54771	Up	Afipia
OTU453	0.318042	215	0.548807	Up	Faecalibacterium

OTU296	0.326377	216	0.560583	Up	Papillibacter
OTU166	0.328441	217	0.561529	Down	Lachnospiraceae Incertae Sedis
OTU7	0.330993	218	0.563296	Up	Bacteroides
OTU256	0.33172	219	0.561955	Up	Anaerotruncus
OTU274	0.333905	220	0.563085	Down	Lachnospiraceae Incertae Sedis
OTU65	0.334251	221	0.561118	Up	Lachnospiraceae Incertae Sedis
OTU327	0.337489	222	0.564002	Up	Pelomonas
OTU168	0.342414	223	0.569666	Down	Roseburia
OTU89	0.347493	224	0.575535	Up	Bacteroides
OTU71	0.353559	225	0.582979	Up	Lachnospiraceae Incertae Sedis
OTU47	0.353621	226	0.580501	Down	Succinispira
OTU349	0.371504	227	0.607171	Up	Syntrophococcus
OTU495	0.372554	228	0.606217	Down	Streptococcus
OTU304	0.375615	229	0.608529	Down	Faecalibacterium
OTU181	0.376974	230	0.608075	Up	Bacteroides
OTU199	0.379331	231	0.609229	Up	Acetanaerobacterium
OTU44	0.383199	232	0.612788	Up	Lachnospiraceae Incertae Sedis
OTU183	0.383518	233	0.610665	Down	Bacteroides
OTU364	0.384954	234	0.610333	Up	Exiguobacterium
OTU6	0.403239	235	0.636604	Down	Lachnospiraceae Incertae Sedis
OTU553	0.403416	236	0.634184	Up	Syntrophococcus
OTU88	0.409553	237	0.641115	Down	Streptococcus
OTU268	0.412992	238	0.643782	Up	Staphylococcus
OTU198	0.417755	239	0.648482	Up	Lachnospiraceae Incertae Sedis
OTU160	0.428286	240	0.662059	Down	Lachnospiraceae Incertae Sedis
OTU315	0.440228	241	0.677696	Down	Coriobacterineae
OTU20	0.44566	242	0.683222	Down	Lachnospiraceae Incertae Sedis
OTU354	0.450531	243	0.687848	Up	Anaerotruncus
OTU179	0.450803	244	0.685442	Up	Ruminococcaceae Incertae Sedis
OTU76	0.454998	245	0.688997	Down	Lachnobacterium
OTU374	0.455869	246	0.687509	Down	Lachnospiraceae Incertae Sedis
OTU4	0.464125	247	0.697128	Up	Lachnospiraceae Incertae Sedis
OTU24	0.466828	248	0.69836	Up	Lachnospiraceae Incertae Sedis
OTU173	0.473245	249	0.705117	Down	Anaerotruncus

OTU54	0.476242	250	0.706743	Up	Lachnospiraceae Incertae Sedis
OTU288	0.477369	251	0.705593	Up	Ruminococcaceae Incertae Sedis
OTU229	0.478121	252	0.703901	Down	Coriobacterineae
OTU367	0.484431	253	0.710371	Up	Pseudomonas
OTU233	0.495265	254	0.723399	Up	Syntrophococcus
OTU359	0.499339	255	0.72649	Up	Faecalibacterium
OTU452	0.505628	256	0.732766	Down	Butyrivibrio
OTU455	0.508508	257	0.734071	Down	Finegoldia
OTU41	0.508672	258	0.731462	Down	Subdoligranulum
OTU62	0.508801	259	0.728823	Down	Ruminococcus
OTU400	0.515068	260	0.734962	Up	Bryantella
OTU42	0.519408	261	0.738315	Up	Prevotella
OTU470	0.521033	262	0.737799	Down	Lachnospiraceae Incertae Sedis
OTU422	0.524664	263	0.740116	Up	Peptococcaceae 1
OTU566	0.531236	264	0.746548	Down	Dorea
OTU214	0.531345	265	0.743883	Down	Roseburia
OTU375	0.534803	266	0.74591	Up	Pseudomonas
OTU456	0.541252	267	0.752076	Down	Anaerovorax
OTU538	0.541252	268	0.74927	Down	Lachnospiraceae Incertae Sedis
OTU272	0.543323	269	0.749342	Down	Sporobacter
OTU182	0.544691	270	0.748446	Down	Lachnospiraceae Incertae Sedis
OTU260	0.549257	271	0.751935	Down	Erysipelotrichaceae Incertae Sedis
OTU406	0.551284	272	0.751935	Up	Bacteroides
OTU17	0.554959	273	0.754175	Down	Escherichia
OTU123	0.562088	274	0.761075	Up	Papillibacter
OTU58	0.577186	275	0.778677	Down	Peptostreptococcaceae Incertae Sedis
OTU380	0.597757	276	0.803507	Down	Sporobacter
OTU372	0.598207	277	0.801208	Up	Allomonas
OTU460	0.598207	278	0.798326	Up	Lachnospiraceae Incertae Sedis
OTU164	0.598254	279	0.795527	Down	Faecalibacterium
OTU9	0.606837	280	0.804058	Up	Bacteroides
OTU493	0.611938	281	0.807932	Down	Lachnospiraceae Incertae Sedis
OTU411	0.61495	282	0.80903	Up	Faecalibacterium
OTU506	0.61495	283	0.806172	Up	Syntrophococcus
OTU104	0.620801	284	0.810976	Down	Syntrophococcus
OTU184	0.621999	285	0.80969	Down	Lachnospiraceae Incertae Sedis

OTU60	0.622167	286	0.807077	Up	Subdoligranulum
OTU196	0.627379	287	0.811003	Down	Bacteroides
OTU305	0.635906	288	0.819171	Down	Lachnospiraceae Incertae Sedis
OTU408	0.636907	289	0.817621	Up	Bryantella
OTU217	0.637392	290	0.815422	Up	Prevotella
OTU27	0.644638	291	0.821858	Up	Lachnospiraceae Incertae Sedis
OTU117	0.644751	292	0.819187	Down	Naxibacter
OTU238	0.648684	293	0.821372	Down	Lachnospiraceae Incertae Sedis
OTU129	0.649316	294	0.819374	Down	Roseburia
OTU148	0.651838	295	0.819769	Down	Lachnospiraceae Incertae Sedis
OTU343	0.668166	296	0.837465	Up	Lachnobacterium
OTU429	0.668166	297	0.834645	Down	Dorea
OTU363	0.670411	298	0.834639	Up	Faecalibacterium
OTU140	0.671784	299	0.833551	Up	Faecalibacterium
OTU52	0.672431	300	0.831573	Up	Lachnospiraceae Incertae Sedis
OTU378	0.689349	301	0.849663	Down	Bacillaceae 1
OTU508	0.689557	302	0.847104	Down	Lachnospiraceae Incertae Sedis
OTU10	0.689926	303	0.844761	Up	Coprobacillus
OTU32	0.690686	304	0.84291	Down	Erysipelotrichaceae Incertae Sedis
OTU80	0.698714	305	0.849911	Down	Lachnospiraceae Incertae Sedis
OTU110	0.712924	306	0.864363	Up	Lachnospiraceae Incertae Sedis
OTU106	0.715991	307	0.865253	Down	Lachnospiraceae Incertae Sedis
OTU379	0.716925	308	0.863568	Up	Roseburia
OTU171	0.716992	309	0.860854	Down	Bacteroides
OTU30	0.725113	310	0.867797	Up	Bryantella
OTU324	0.738903	311	0.881456	Up	Faecalibacterium
OTU311	0.740828	312	0.880921	Up	Lachnospiraceae Incertae Sedis
OTU101	0.745441	313	0.883574	Down	Pseudoalteromonas
OTU287	0.751988	314	0.888496	Down	Anaerovorax
OTU212	0.757145	315	0.891749	Down	Coprobacillus
OTU55	0.767222	316	0.900757	Up	Parabacteroides
OTU392	0.768645	317	0.899582	Up	Lachnospiraceae Incertae Sedis
OTU114	0.768686	318	0.8968	Up	Megamonas
OTU243	0.772843	319	0.898824	Up	Anaerotruncus

OTU108	0.77323	320	0.896464	Up	Lachnospiraceae Incertae Sedis
OTU231	0.775025	321	0.895745	Up	Anaerotruncus
OTU316	0.775025	322	0.892964	Up	Alistipes
OTU403	0.784314	323	0.900868	Up	Methylobacterium
OTU131	0.784488	324	0.898287	Up	Lachnospiraceae Incertae Sedis
OTU103	0.789604	325	0.901363	Up	Roseburia
OTU105	0.793064	326	0.902536	Up	Bacteroides
OTU155	0.800433	327	0.908137	Down	Roseburia
OTU107	0.811899	328	0.918337	Down	Ruminococcus
OTU269	0.815747	329	0.919885	Down	Butyrivibrio
OTU312	0.819071	330	0.920834	Down	Coriobacterineae
OTU18	0.822123	331	0.921474	Up	Faecalibacterium
OTU115	0.825146	332	0.922076	Down	Roseburia
OTU126	0.825636	333	0.919852	Down	Aeromonas
OTU40	0.830942	334	0.922993	Up	Lachnospiraceae Incertae Sedis
OTU12	0.832163	335	0.921589	Up	Bryantella
OTU416	0.838341	336	0.925668	Up	Lachnospiraceae Incertae Sedis
OTU102	0.839205	337	0.923873	Down	Lachnospiraceae Incertae Sedis
OTU130	0.847691	338	0.930453	Up	Lachnospiraceae Incertae Sedis
OTU51	0.849066	339	0.929213	Down	Klebsiella
OTU187	0.853675	340	0.93151	Down	Erysipelotrichaceae Incertae Sedis
OTU492	0.860391	341	0.936085	Down	Coriobacterineae
OTU158	0.870215	342	0.944005	Down	Bacteroides
OTU43	0.871472	343	0.942613	Down	Lachnospiraceae Incertae Sedis
OTU445	0.874152	344	0.942763	Down	Corynebacterineae
OTU424	0.874975	345	0.940915	Down	Streptococcus
OTU35	0.885406	346	0.949381	Down	Bryantella
OTU358	0.886366	347	0.947671	Up	Roseburia
OTU39	0.889892	348	0.948707	Down	Coriobacterineae
OTU291	0.890838	349	0.946994	Up	Syntrophococcus
OTU292	0.892843	350	0.946414	Down	Alistipes
OTU94	0.894124	351	0.945072	Down	Anaerotruncus
OTU31	0.903421	352	0.952185	Up	Coprococcus
OTU399	0.913216	353	0.959782	Down	Ralstonia
OTU253	0.914073	354	0.957969	Down	Uruburuella
OTU69	0.921491	355	0.963023	Down	Lachnospiraceae Incertae Sedis

OTU547	0.921893	356	0.960737	Up	Subdoligranulum
OTU25	0.931086	357	0.967599	Up	Parabacteroides
OTU277	0.933541	358	0.967441	Down	Lachnospiraceae Incertae Sedis
OTU293	0.935543	359	0.966814	Down	Lachnospiraceae Incertae Sedis
OTU98	0.93936	360	0.968063	Up	Lachnospiraceae Incertae Sedis
OTU194	0.949283	361	0.975579	Down	Alistipes
OTU344	0.961288	362	0.985187	Down	Carnobacteriaceae 1
OTU48	0.967805	363	0.989134	Down	Bacteroides
OTU132	0.972304	364	0.991002	Down	Parabacteroides
OTU355	0.973371	365	0.989371	Down	Corynebacterineae
OTU458	0.984021	366	0.997463	Up	Roseburia
OTU180	0.98511	367	0.995847	Down	Roseburia
OTU151	0.985591	368	0.993626	Down	Subdoligranulum
OTU16	0.986197	369	0.991542	Down	Lachnospiraceae Incertae Sedis
OTU2	0.986203	370	0.988868	Up	Faecalibacterium
OTU150	0.995379	371	0.995379	Up	Ruminococcaceae Incertae Sedis

Supplementary Table 4: Kruskal-Wallis tests on log-normalized abundances of OTUs (97%) in BMI categories Normal (<25) vs. Overweight (26- 30) vs. Obese (>30). RDP classification of consensus sequences at genus level shown. Only OTUs which have at least 1 sequence assigned to them in 25% of the samples are shown. Kruskal-Wallis p-Values were corrected for multiple testing[1] using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted Rank of the taxon.

OTUname	Kruskal-Wallis p-Value	Rank	n*p/R	RDP Genus level Assignment
OTU153	0.0125	1	4.6375	Roseburia
OTU306	0.0202	2	3.7471	Oligotropha
OTU445	0.0252	3	3.1164	Corynebacterineae
OTU4	0.0256	4	2.3744	Lachnospiraceae Incertae Sedis
OTU538	0.0295	5	2.1889	Lachnospiraceae Incertae Sedis
OTU439	0.037	6	2.28783	Algibacter
OTU72	0.0371	7	1.9663	Aquabacterium
OTU525	0.0374	8	1.73443	Catonella
OTU75	0.0376	9	1.54996	Stenotrophomonas
OTU110	0.0412	10	1.52852	Lachnospiraceae Incertae Sedis
OTU98	0.0416	11	1.40305	Lachnospiraceae Incertae Sedis
OTU277	0.0429	12	1.32633	Lachnospiraceae Incertae Sedis

OTU28	0.0442	13	1.2614	Bacteroides
OTU156	0.0452	14	1.1978	Lachnospiraceae Incertae Sedis
OTU16	0.0517	15	1.27871	Lachnospiraceae Incertae Sedis
OTU43	0.054	16	1.25213	Lachnospiraceae Incertae Sedis
OTU27	0.0549	17	1.19811	Lachnospiraceae Incertae Sedis
OTU470	0.0686	18	1.41392	Lachnospiraceae Incertae Sedis
OTU39	0.0705	19	1.37661	Coriobacterineae
OTU506	0.0736	20	1.36528	Syntrophococcus
OTU157	0.0758	21	1.33913	Marinilabilia
OTU9	0.0786	22	1.32548	Bacteroides
OTU131	0.0788	23	1.27108	Lachnospiraceae Incertae Sedis
OTU240	0.0798	24	1.23358	Weissella
OTU566	0.0815	25	1.20946	Dorea
OTU288	0.0848	26	1.21003	Ruminococcaceae Incertae Sedis
OTU1	0.0869	27	1.19407	Bacteroides
OTU341	0.0879	28	1.16468	Prevotella
OTU326	0.0911	29	1.16545	Lachnospiraceae Incertae Sedis
OTU380	0.0947	30	1.17112	Sporobacter
OTU214	0.0954	31	1.14172	Roseburia
OTU11	0.0984	32	1.14083	Bacteroides
OTU172	0.0997	33	1.12087	Marinilabilia
OTU173	0.1008	34	1.09991	Anaerotruncus
OTU499	0.1021	35	1.08226	Lachnospiraceae Incertae Sedis
OTU7	0.1026	36	1.05735	Bacteroides
OTU357	0.1084	37	1.08693	Coprococcus
OTU356	0.1086	38	1.06028	Novosphingobium
OTU248	0.1124	39	1.06924	Lachnospiraceae Incertae Sedis
OTU328	0.1146	40	1.06292	Parasporobacterium
OTU56	0.119	41	1.0768	Delftia
OTU96	0.1197	42	1.05735	Diaphorobacter
OTU372	0.1223	43	1.05519	Allomonas
OTU241	0.1272	44	1.07253	Chryseobacterium
OTU371	0.1295	45	1.06766	Comamonas
OTU305	0.1297	46	1.04606	Lachnospiraceae Incertae Sedis
OTU47	0.1317	47	1.03959	Succinispira
OTU204	0.1363	48	1.05349	Dialister
OTU59	0.1363	49	1.03199	Acinetobacter

OTU138	0.147	50	1.09074	Simkania
OTU519	0.1476	51	1.07372	Catonella
OTU197	0.1479	52	1.05521	Lactobacillus
OTU132	0.1487	53	1.0409	Parabacteroides
OTU79	0.1491	54	1.02437	Lachnospiraceae Incertae Sedis
OTU370	0.1519	55	1.02463	Lactobacillus
OTU97	0.152	56	1.007	Pseudomonas
OTU501	0.1567	57	1.01992	Ruminococcaceae Incertae Sedis
OTU329	0.1616	58	1.03368	Methanohalobium
OTU266	0.1618	59	1.01742	Bacteroides
OTU464	0.1618	60	1.00046	Marinilabilia
OTU338	0.1692	61	1.02907	Micrococcineae
OTU304	0.1731	62	1.03581	Faecalibacterium
OTU374	0.1784	63	1.05058	Lachnospiraceae Incertae Sedis
OTU411	0.1827	64	1.05909	Faecalibacterium
OTU139	0.1839	65	1.04964	Azonexus
OTU399	0.1849	66	1.03936	Ralstonia
OTU40	0.1864	67	1.03216	Lachnospiraceae Incertae Sedis
OTU200	0.1891	68	1.03171	Helicobacter
OTU12	0.1918	69	1.03127	Bryantella
OTU432	0.1919	70	1.01707	Paludibacter
OTU452	0.1938	71	1.01267	Butyrivibrio
OTU86	0.1953	72	1.00634	Fusobacterium
OTU547	0.1959	73	0.9956	Subdoligranulum
OTU51	0.1975	74	0.99017	Klebsiella
OTU148	0.1994	75	0.98637	Lachnospiraceae Incertae Sedis
OTU391	0.2026	76	0.98901	Aquiflexum
OTU120	0.2027	77	0.97665	Micrococcineae
OTU367	0.2053	78	0.97649	Pseudomonas
OTU287	0.2077	79	0.9754	Anaerovorax
OTU412	0.2092	80	0.97017	Sphingomonas
OTU502	0.2095	81	0.95956	Paludibacter
OTU319	0.2113	82	0.956	Agrobacterium
OTU23	0.215	83	0.96102	Lachnospiraceae Incertae Sedis
OTU269	0.2155	84	0.95179	Butyrivibrio
OTU177	0.2167	85	0.94583	Butyrivibrio
OTU437	0.2182	86	0.9413	Marinilabilia

OTU136	0.2206	87	0.94072	Micrococcineae
OTU182	0.2221	88	0.93635	Lachnospiraceae Incertae Sedis
OTU243	0.223	89	0.92958	Anaerotruncus
OTU14	0.2291	90	0.9444	Erysipelotrichaceae Incertae Sedis
OTU283	0.2296	91	0.93606	Anaerophaga
OTU421	0.2297	92	0.92629	Streptococcus
OTU238	0.2308	93	0.92072	Lachnospiraceae Incertae Sedis
OTU442	0.2308	94	0.91092	Roseburia
OTU492	0.2332	95	0.91071	Coriobacterineae
OTU29	0.235	96	0.90818	Lachnospiraceae Incertae Sedis
OTU406	0.2368	97	0.9057	Bacteroides
OTU265	0.2376	98	0.89949	Sphingomonas
OTU90	0.2431	99	0.91101	Lachnospiraceae Incertae Sedis
OTU38	0.2507	100	0.9301	Pseudomonas
OTU32	0.251	101	0.92199	Erysipelotrichaceae Incertae Sedis
OTU458	0.2529	102	0.91986	Roseburia
OTU474	0.2555	103	0.9203	Sphingobium
OTU569	0.259	104	0.92393	Erwinia
OTU101	0.2611	105	0.92255	Pseudoalteromonas
OTU162	0.2672	106	0.9352	Veillonella
OTU22	0.2693	107	0.93374	Acidovorax
OTU37	0.2702	108	0.92819	Cloacibacterium
OTU416	0.2715	109	0.9241	Lachnospiraceae Incertae Sedis
OTU80	0.273	110	0.92075	Lachnospiraceae Incertae Sedis
OTU392	0.2753	111	0.92015	Lachnospiraceae Incertae Sedis
OTU87	0.2765	112	0.91591	Propionibacterineae
OTU161	0.2781	113	0.91305	Prevotella
OTU109	0.2825	114	0.91936	Turicibacter
OTU297	0.2949	115	0.95137	Bacillaceae 1
OTU216	0.3	116	0.95948	Sphingomonas
OTU127	0.3011	117	0.95477	Lachnospiraceae Incertae Sedis
OTU256	0.3017	118	0.94857	Anaerotruncus
OTU195	0.3058	119	0.95338	Pseudoalteromonas
OTU119	0.3065	120	0.9476	Lachnobacterium
OTU239	0.3065	121	0.93976	Succinispira
OTU183	0.3107	122	0.94483	Bacteroides

OTU146	0.3111	123	0.93836	Vibrio
OTU70	0.3138	124	0.93887	Sphingobium
OTU300	0.3145	125	0.93344	Lachnospiraceae Incertae Sedis
OTU354	0.3245	126	0.95547	Anaerotruncus
OTU128	0.3258	127	0.95175	Prevotella
OTU345	0.3295	128	0.95504	Butyrivibrio
OTU144	0.3315	129	0.95338	Dorea
OTU133	0.3389	130	0.96717	Faecalibacterium
OTU393	0.3441	131	0.97451	Micrococcineae
OTU401	0.3465	132	0.97388	Alistipes
OTU226	0.3468	133	0.96739	Rikenella
OTU313	0.347	134	0.96072	Enterobacter
OTU454	0.3474	135	0.95471	Paludibacter
OTU6	0.3478	136	0.94878	Lachnospiraceae Incertae Sedis
OTU118	0.3482	137	0.94294	Burkholderia
OTU176	0.3533	138	0.94981	Erwinia
OTU397	0.357	139	0.95286	Peptostreptococcaceae Incertae Sedis
OTU180	0.3577	140	0.94791	Roseburia
OTU168	0.3627	141	0.95434	Roseburia
OTU419	0.3647	142	0.95284	Micrococcineae
OTU50	0.3647	143	0.94618	Sutterella
OTU34	0.3652	144	0.9409	Dorea
OTU71	0.3653	145	0.93466	Lachnospiraceae Incertae Sedis
OTU64	0.3681	146	0.93538	Erwinia
OTU159	0.375	147	0.94643	Faecalibacterium
OTU199	0.376	148	0.94254	Acetanaerobacterium
OTU88	0.3762	149	0.93671	Streptococcus
OTU178	0.3777	150	0.93418	Lachnospiraceae Incertae Sedis
OTU352	0.3778	151	0.92824	Saprospira
OTU237	0.381	152	0.92994	Prevotella
OTU210	0.3815	153	0.92508	Allobaculum
OTU225	0.3842	154	0.92557	Prevotella
OTU74	0.3866	155	0.92535	Ruminococcus
OTU334	0.3908	156	0.9294	Citrobacter
OTU192	0.3917	157	0.92561	Sphingomonas
OTU158	0.3954	158	0.92844	Bacteroides
OTU353	0.396	159	0.924	Dorea

OTU229	0.4	160	0.9275	Coriobacterineae
OTU193	0.4004	161	0.92266	Xylanibacter
OTU230	0.4021	162	0.92086	Butyrivibrio
OTU57	0.4051	163	0.92204	Lachnospiraceae Incertae Sedis
OTU19	0.409	164	0.92524	Syntrophococcus
OTU363	0.4092	165	0.92008	Faecalibacterium
OTU65	0.4105	166	0.91744	Lachnospiraceae Incertae Sedis
OTU145	0.4157	167	0.9235	Afipia
OTU270	0.4187	168	0.92463	Succinispira
OTU84	0.4201	169	0.92223	Marinomonas
OTU100	0.4225	170	0.92204	Xylanibacter
OTU366	0.4227	171	0.91709	Coprococcus
OTU403	0.4238	172	0.91413	Methylobacterium
OTU267	0.4253	173	0.91206	Parabacteroides
OTU170	0.4256	174	0.90746	Bacteroides
OTU423	0.43	175	0.9116	Parasporobacterium
OTU268	0.4307	176	0.9079	Staphylococcus
OTU365	0.4311	177	0.90361	Succinispira
OTU181	0.4312	178	0.89874	Bacteroides
OTU364	0.4323	179	0.896	Exiguobacterium
OTU491	0.4335	180	0.89349	Clostridiaceae 1
OTU105	0.4364	181	0.8945	Bacteroides
OTU5	0.4368	182	0.8904	Sphingomonas
OTU322	0.4414	183	0.89486	Roseburia
OTU224	0.4432	184	0.89363	Prevotella
OTU213	0.4468	185	0.89602	Lactococcus
OTU343	0.4495	186	0.89658	Lachnobacterium
OTU26	0.4516	187	0.89596	Dorea
OTU49	0.4579	188	0.90362	Sutterella
OTU186	0.4584	189	0.89982	Faecalibacterium
OTU45	0.4603	190	0.8988	Xenohaliotis
OTU344	0.4722	191	0.91721	Carnobacteriaceae 1
OTU114	0.4744	192	0.91668	Megamonas
OTU194	0.478	193	0.91885	Alistipes
OTU249	0.4809	194	0.91966	Faecalibacterium
OTU73	0.4888	195	0.92997	Lactococcus
OTU122	0.4898	196	0.92712	Prevotella

OTU307	0.4912	197	0.92505	Megamonas
OTU124	0.5009	198	0.93856	Lactobacillus
OTU187	0.5039	199	0.93943	Erysipelotrichaceae Incertae Sedis
OTU235	0.5047	200	0.93622	Desulfovibrio
OTU149	0.5059	201	0.93378	Haemophilus
OTU309	0.5061	202	0.92952	Paludibacter
OTU143	0.5074	203	0.92732	Lachnospiraceae Incertae Sedis
OTU31	0.5076	204	0.92314	Coprococcus
OTU30	0.5115	205	0.92569	Bryantella
OTU151	0.5116	206	0.92138	Subdoligranulum
OTU425	0.5166	207	0.92589	Enhydrobacter
OTU41	0.5176	208	0.92322	Subdoligranulum
OTU291	0.5193	209	0.92182	Syntrophococcus
OTU82	0.5226	210	0.92326	Roseburia
OTU206	0.5229	211	0.91941	Paludibacter
OTU160	0.5232	212	0.9156	Lachnospiraceae Incertae Sedis
OTU135	0.5243	213	0.91322	Clostridiaceae 1
OTU418	0.5253	214	0.91068	Stenotrophomonas
OTU152	0.5303	215	0.91508	Faecalibacterium
OTU46	0.5305	216	0.91118	Bacillaceae 1
OTU76	0.5306	217	0.90715	Lachnobacterium
OTU89	0.5315	218	0.90453	Bacteroides
OTU330	0.532	219	0.90124	Coriobacterineae
OTU471	0.535	220	0.9022	Lachnospiraceae Incertae Sedis
OTU171	0.5368	221	0.90114	Bacteroides
OTU103	0.5438	222	0.90878	Roseburia
OTU244	0.5447	223	0.9062	Prevotella
OTU358	0.5453	224	0.90315	Roseburia
OTU453	0.5461	225	0.90046	Faecalibacterium
OTU111	0.5483	226	0.90009	Peptostreptococcaceae Incertae Sedis
OTU189	0.5493	227	0.89775	Acidovorax
OTU24	0.55	228	0.89496	Lachnospiraceae Incertae Sedis
OTU376	0.5502	229	0.89137	Methylobacterium
OTU203	0.5533	230	0.8925	Rheinheimera
OTU455	0.5625	231	0.90341	Finegoldia
OTU484	0.5693	232	0.91039	Effluviibacter

OTU350	0.5747	233	0.91508	Coprococcus
OTU35	0.5757	234	0.91276	Bryantella
OTU69	0.5784	235	0.91313	Lachnospiraceae Incertae Sedis
OTU91	0.5813	236	0.91382	Lactobacillus
OTU66	0.5835	237	0.91341	Streptococcus
OTU463	0.5846	238	0.91129	Lachnospiraceae Incertae Sedis
OTU387	0.58818	239	0.91303	Coprococcus
OTU378	0.589	240	0.9105	Bacillaceae 1
OTU126	0.5937	241	0.91395	Aeromonas
OTU373	0.5949	242	0.91202	Sporobacter
OTU169	0.595	243	0.90842	Streptococcus
OTU233	0.5959	244	0.90606	Syntrophococcus
OTU284	0.5973	245	0.90448	Rubritepida
OTU108	0.6038	246	0.91061	Lachnospiraceae Incertae Sedis
OTU247	0.6044	247	0.90782	Xylanibacter
OTU130	0.6073	248	0.9085	Lachnospiraceae Incertae Sedis
OTU165	0.6145	249	0.91558	Alistipes
OTU327	0.615	250	0.91266	Pelomonas
OTU106	0.6165	251	0.91124	Lachnospiraceae Incertae Sedis
OTU420	0.6168	252	0.90807	Dorea
OTU207	0.6187	253	0.90726	Succinispira
OTU324	0.6203	254	0.90603	Faecalibacterium
OTU275	0.6213	255	0.90393	Lachnospiraceae Incertae Sedis
OTU347	0.6235	256	0.90359	Vitellibacter
OTU198	0.6266	257	0.90455	Lachnospiraceae Incertae Sedis
OTU493	0.6268	258	0.90133	Lachnospiraceae Incertae Sedis
OTU60	0.6291	259	0.90114	Subdoligranulum
OTU164	0.6307	260	0.89996	Faecalibacterium
OTU85	0.6349	261	0.90248	Bacteroides
OTU155	0.6395	262	0.90555	Roseburia
OTU188	0.6396	263	0.90225	Lachnospiraceae Incertae Sedis
OTU117	0.6399	264	0.89925	Naxibacter
OTU404	0.6453	265	0.90342	Hallella
OTU53	0.6509	266	0.90783	Succinivibrio
OTU67	0.6584	267	0.91486	Lactobacillus
OTU134	0.6601	268	0.9138	Ruminococcaceae Incertae Sedis
OTU286	0.6604	269	0.91081	Hallella

OTU476	0.6642	270	0.91266	Streptococcus
OTU508	0.6654	271	0.91094	Lachnospiraceae Incertae Sedis
OTU361	0.6727	272	0.91754	Succinivibrio
OTU274	0.681	273	0.92546	Lachnospiraceae Incertae Sedis
OTU113	0.6855	274	0.92818	Rikenella
OTU212	0.6881	275	0.92831	Coprobacillus
OTU52	0.69227	276	0.93055	Lachnospiraceae Incertae Sedis
OTU299	0.6954	277	0.93138	Lachnospiraceae Incertae Sedis
OTU315	0.6976	278	0.93097	Coriobacterineae
OTU429	0.6982	279	0.92843	Dorea
OTU107	0.6991	280	0.92631	Ruminococcus
OTU42	0.7035	281	0.92882	Prevotella
OTU20	0.7054	282	0.92803	Lachnospiraceae Incertae Sedis
OTU15	0.7074	283	0.92737	Roseburia
OTU285	0.7114	284	0.92933	Butyrivibrio
OTU102	0.7156	285	0.93154	Lachnospiraceae Incertae Sedis
OTU375	0.7256	286	0.94125	Pseudomonas
OTU389	0.7273	287	0.94017	Parabacteroides
OTU202	0.7275	288	0.93716	Lachnospiraceae Incertae Sedis
OTU222	0.7295	289	0.93649	Prevotella
OTU395	0.7357	290	0.94119	Subdoligranulum
OTU250	0.7363	291	0.93872	Paludibacter
OTU115	0.7405	292	0.94084	Roseburia
OTU21	0.7508	293	0.95067	Finegoldia
OTU33	0.7525	294	0.94958	Lachnospiraceae Incertae Sedis
OTU360	0.7528	295	0.94674	Faecalibacterium
OTU231	0.7545	296	0.94567	Anaerotruncus
OTU292	0.7554	297	0.94361	Alistipes
OTU242	0.7656	298	0.95315	Coriobacterineae
OTU311	0.7664	299	0.95095	Lachnospiraceae Incertae Sedis
OTU205	0.7694	300	0.95149	Erysipelotrichaceae Incertae Sedis
OTU217	0.7694	301	0.94833	Prevotella
OTU140	0.77	302	0.94593	Faecalibacterium
OTU317	0.7757	303	0.94978	Prevotella
OTU190	0.7768	304	0.948	Ruminococcaceae Incertae Sedis
OTU282	0.7852	305	0.95511	Streptococcus
OTU312	0.7899	306	0.95769	Coriobacterineae

OTU303	0.798	307	0.96436	Faecalibacterium
OTU296	0.8006	308	0.96436	Papillibacter
OTU150	0.8055	309	0.96712	Ruminococcaceae Incertae Sedis
OTU184	0.8057	310	0.96424	Lachnospiraceae Incertae Sedis
OTU104	0.8059	311	0.96138	Syntrophococcus
OTU154	0.808	312	0.96079	Faecalibacterium
OTU553	0.8125	313	0.96306	Syntrophococcus
OTU254	0.8131	314	0.9607	Lachnospiraceae Incertae Sedis
OTU359	0.8214	315	0.96743	Faecalibacterium
OTU166	0.8253	316	0.96894	Lachnospiraceae Incertae Sedis
OTU142	0.8254	317	0.966	Lachnospiraceae Incertae Sedis
OTU417	0.8299	318	0.96822	Lachnobacterium
OTU10	0.833	319	0.96879	Coprobacillus
OTU18	0.837	320	0.9704	Faecalibacterium
OTU68	0.8376	321	0.96807	Dorea
OTU3	0.8382	322	0.96575	Lachnospiraceae Incertae Sedis
OTU407	0.839	323	0.96368	Turicibacter
OTU495	0.8404	324	0.96231	Streptococcus
OTU61	0.8405	325	0.95946	Papillibacter
OTU17	0.846	326	0.96278	Escherichia
OTU83	0.8462	327	0.96006	Dorea
OTU54	0.8468	328	0.95781	Lachnospiraceae Incertae Sedis
OTU409	0.848	329	0.95626	Alkalilimnicola
OTU25	0.8491	330	0.95459	Parabacteroides
OTU253	0.8496	331	0.95227	Uruburuella
OTU355	0.8553	332	0.95577	Corynebacterineae
OTU264	0.8585	333	0.95647	Comamonas
OTU129	0.8632	334	0.95882	Roseburia
OTU94	0.8638	335	0.95663	Anaerotruncus
OTU227	0.868	336	0.95842	Lachnospiraceae Incertae Sedis
OTU413	0.8732	337	0.9613	Subdoligranulum
OTU8	0.8757	338	0.9612	Dorea
OTU92	0.8801	339	0.96318	Rubrobacterineae
OTU36	0.8815	340	0.96187	Bacteroides
OTU191	0.8823	341	0.95992	Subdoligranulum
OTU422	0.8834	342	0.95831	Peptococcaceae 1
OTU396	0.8849	343	0.95714	Coprococcus

OTU167	0.8882	344	0.95791	Allobaculum
OTU93	0.895	345	0.96245	Alistipes
OTU408	0.8976	346	0.96246	Bryantella
OTU260	0.9	347	0.96225	Erysipelotrichaceae Incertae Sedis
OTU2	0.9165	348	0.97707	Faecalibacterium
OTU456	0.9187	349	0.97661	Anaerovorax
OTU293	0.9214	350	0.97668	Lachnospiraceae Incertae Sedis
OTU219	0.9222	351	0.97475	Rikenella
OTU349	0.9245	352	0.9744	Syntrophococcus
OTU460	0.9246	353	0.97175	Lachnospiraceae Incertae Sedis
OTU95	0.9326	354	0.97739	Ruminococcus
OTU48	0.9459	355	0.98853	Bacteroides
OTU55	0.9609	356	1.00139	Parabacteroides
OTU196	0.9689	357	1.0069	Bacteroides
OTU368	0.9705	358	1.00574	Ruminococcaceae Incertae Sedis
OTU424	0.9713	359	1.00377	Streptococcus
OTU137	0.9718	360	1.00149	Prevotella
OTU123	0.9789	361	1.00602	Papillibacter
OTU316	0.9789	362	1.00324	Alistipes
OTU62	0.9824	363	1.00405	Ruminococcus
OTU272	0.9832	364	1.00211	Sporobacter
OTU379	0.9862	365	1.00241	Roseburia
OTU44	0.9892	366	1.00271	Lachnospiraceae Incertae Sedis
OTU141	0.9895	367	1.00028	Faecalibacterium
OTU58	0.9913	368	0.99938	Peptostreptococcaceae Incertae Sedis
OTU400	0.9926	369	0.99798	Bryantella
OTU179	0.9933	370	0.99598	Ruminococcaceae Incertae Sedis
OTU77	0.9993	371	0.9993	Coprococcus

Supplementary Table 5: Kruskal Wallis-tests on log-normalized abundances of OTUs (97%) in WHR levels low, medium and high. Only OTUs which have at least 1 sequence assigned to them in 25% of the samples are shown. RDP classification of consensus sequences at genus level shown. Kruskal-Wallis p-Values were corrected for multiple testing[1] using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted rank of the taxon.

OTUName	Kruskal-Wallis p-Value	Rank	n*p/R	RDP Genus Level Assignment
OTU299	0.0059	1	2.1889	Lachnospiraceae Incertae Sedis
OTU538	0.0068	2	1.2614	Lachnospiraceae Incertae Sedis
OTU306	0.0149	3	1.84263	Oligotropha
OTU569	0.0174	4	1.61385	Erwinia
OTU387	0.022	5	1.6324	Coprococcus
OTU349	0.0265	6	1.63858	Syntrophococcus
OTU8	0.0268	7	1.4204	Dorea
OTU419	0.0338	8	1.56748	Micrococcineae
OTU484	0.0349	9	1.43866	Effluviibacter
OTU19	0.0404	10	1.49884	Syntrophococcus
OTU464	0.0406	11	1.36933	Marinilabilia
OTU156	0.0414	12	1.27995	Lachnospiraceae Incertae Sedis
OTU248	0.0432	13	1.23286	Lachnospiraceae Incertae Sedis
OTU48	0.046	14	1.219	Bacteroides
OTU210	0.0463	15	1.14515	Allobaculum
OTU172	0.048	16	1.113	Marinilabilia
OTU93	0.0497	17	1.08463	Alistipes
OTU373	0.0556	18	1.14598	Sporobacter
OTU168	0.0571	19	1.11495	Roseburia
OTU250	0.0588	20	1.09074	Paludibacter
OTU375	0.0613	21	1.08297	Pseudomonas
OTU291	0.0616	22	1.0388	Syntrophococcus
OTU35	0.0698	23	1.1259	Bryantella
OTU357	0.0708	24	1.09445	Coprococcus
OTU439	0.071	25	1.05364	Algibacter
OTU110	0.0715	26	1.02025	Lachnospiraceae Incertae Sedis
OTU525	0.0717	27	0.98521	Catonella
OTU67	0.0736	28	0.9752	Lactobacillus

OTU5	0.0741	29	0.94797	Sphingomonas
OTU96	0.0766	30	0.94729	Diaphorobacter
OTU493	0.0787	31	0.94186	Lachnospiraceae Incertae Sedis
OTU566	0.0835	32	0.96808	Dorea
OTU84	0.0839	33	0.94324	Marinomonas
OTU34	0.0849	34	0.92641	Dorea
OTU399	0.0853	35	0.90418	Ralstonia
OTU366	0.0882	36	0.90895	Coprococcus
OTU142	0.0913	37	0.91547	Lachnospiraceae Incertae Sedis
OTU95	0.0916	38	0.89431	Ruminococcus
OTU360	0.0918	39	0.87328	Faecalibacterium
OTU45	0.0918	40	0.85145	Xenohaliotis
OTU508	0.0926	41	0.83792	Lachnospiraceae Incertae Sedis
OTU329	0.0961	42	0.84888	Methanohalobium
OTU151	0.0962	43	0.83	Subdoligranulum
OTU501	0.0979	44	0.82548	Ruminococcaceae Incertae Sedis
OTU244	0.1002	45	0.82609	Prevotella
OTU315	0.1064	46	0.85814	Coriobacterineae
OTU553	0.1072	47	0.8462	Syntrophococcus
OTU230	0.1095	48	0.84634	Butyrivibrio
OTU316	0.1102	49	0.83437	Alistipes
OTU197	0.1107	50	0.82139	Lactobacillus
OTU104	0.1147	51	0.83439	Syntrophococcus
OTU191	0.1181	52	0.8426	Subdoligranulum
OTU161	0.1184	53	0.8288	Prevotella
OTU243	0.1184	54	0.81345	Anaerotruncus
OTU62	0.1192	55	0.80406	Ruminococcus
OTU23	0.1193	56	0.79036	Lachnospiraceae Incertae Sedis
OTU205	0.1197	57	0.7791	Erysipelotrichaceae Incertae Sedis
OTU106	0.125	58	0.79957	Lachnospiraceae Incertae Sedis
OTU224	0.1271	59	0.79922	Prevotella
OTU74	0.131	60	0.81002	Ruminococcus
OTU372	0.1312	61	0.79795	Allomonas
OTU470	0.1338	62	0.80064	Lachnospiraceae Incertae Sedis

OTU160	0.1368	63	0.8056	Lachnospiraceae Incertae Sedis
OTU404	0.1385	64	0.80287	Hallella
OTU190	0.1394	65	0.79565	Ruminococcaceae Incertae Sedis
OTU432	0.1402	66	0.78809	Paludibacter
OTU471	0.1412	67	0.78187	Lachnospiraceae Incertae Sedis
OTU28	0.144	68	0.78565	Bacteroides
OTU233	0.145	69	0.77964	Syntrophococcus
OTU41	0.1468	70	0.77804	Subdoligranulum
OTU365	0.1534	71	0.80157	Succinispira
OTU395	0.1557	72	0.80229	Subdoligranulum
OTU305	0.1573	73	0.79943	Lachnospiraceae Incertae Sedis
OTU30	0.1594	74	0.79915	Bryantella
OTU154	0.1597	75	0.78998	Faecalibacterium
OTU46	0.1602	76	0.78203	Bacillaceae 1
OTU100	0.1611	77	0.77621	Xylanibacter
OTU254	0.1671	78	0.7948	Lachnospiraceae Incertae Sedis
OTU200	0.1725	79	0.81009	Helicobacter
OTU421	0.1763	80	0.81759	Streptococcus
OTU277	0.1773	81	0.81208	Lachnospiraceae Incertae Sedis
OTU239	0.1778	82	0.80444	Succinispira
OTU1	0.1808	83	0.80815	Bacteroides
OTU68	0.1814	84	0.80118	Dorea
OTU72	0.1816	85	0.79263	Aquabacterium
OTU495	0.1891	86	0.81577	Streptococcus
OTU275	0.1938	87	0.82643	Lachnospiraceae Incertae Sedis
OTU370	0.1946	88	0.82042	Lactobacillus
OTU284	0.1958	89	0.8162	Rubritepida
OTU195	0.1959	90	0.80754	Pseudoalteromonas
OTU91	0.1979	91	0.80682	Lactobacillus
OTU82	0.198	92	0.79846	Roseburia
OTU378	0.1982	93	0.79067	Bacillaceae 1
OTU206	0.2061	94	0.81344	Paludibacter
OTU317	0.2063	95	0.80566	Prevotella
OTU165	0.2065	96	0.79804	Alistipes
OTU113	0.2074	97	0.79325	Rikenella

OTU130	0.2101	98	0.79538	Lachnospiraceae Incertae Sedis
OTU138	0.2157	99	0.80833	Acidovorax
OTU22	0.2166	100	0.80359	Coriobacterineae
OTU492	0.2189	101	0.80408	Lactococcus
OTU73	0.2211	102	0.8042	Prevotella
OTU137	0.225	103	0.81044	Afipia
OTU145	0.23	104	0.82048	Erwinia
OTU64	0.2302	105	0.81337	Streptococcus
OTU282	0.2306	106	0.8071	Prevotella
OTU42	0.231	107	0.80094	Enhydrobacter
OTU425	0.2351	108	0.80761	Cloacibacterium
OTU37	0.2366	109	0.80531	Papillibacter
OTU61	0.2382	110	0.80338	Roseburia
OTU180	0.2389	111	0.79849	Streptococcus
OTU169	0.2395	112	0.79334	Micrococcineae
OTU136	0.2416	113	0.79322	Faecalibacterium
OTU304	0.2444	114	0.79537	Lachnospiraceae Incertae Sedis
OTU188	0.2467	115	0.79588	Coprobacillus
OTU10	0.2477	116	0.79221	Prevotella
OTU128	0.2568	117	0.8143	Dorea
OTU420	0.2582	118	0.8118	Paludibacter
OTU454	0.2585	119	0.80591	Uruburuella
OTU253	0.2599	120	0.80352	Bacteroides
OTU406	0.2601	121	0.7975	Bacteroides
OTU7	0.2613	122	0.79461	Weissella
OTU240	0.2614	123	0.78845	Coriobacterineae
OTU312	0.2621	124	0.78419	Acinetobacter
OTU59	0.2645	125	0.78504	Acidovorax
OTU189	0.2663	126	0.78411	Rubrobacterineae
OTU92	0.2691	127	0.78611	Xylanibacter
OTU193	0.2737	128	0.7933	Streptococcus
OTU424	0.2749	129	0.7906	Papillibacter
OTU123	0.2753	130	0.78566	Ruminococcaceae Incertae Sedis
OTU368	0.2773	131	0.78533	Faecalibacterium
OTU18	0.2803	132	0.78781	Bryantella

OTU12	0.2818	133	0.78607	Sphingomonas
OTU192	0.284	134	0.7863	Succinispira
OTU207	0.284	135	0.78047	Lachnospiraceae Incertae Sedis
OTU416	0.2856	136	0.7791	Allobaculum
OTU167	0.2875	137	0.77856	Lachnospiraceae Incertae Sedis
OTU98	0.2908	138	0.78179	Faecalibacterium
OTU249	0.2916	139	0.7783	Lachnospiraceae Incertae Sedis
OTU300	0.2948	140	0.78122	Roseburia
OTU214	0.2976	141	0.78305	Klebsiella
OTU51	0.299	142	0.78119	Streptococcus
OTU476	0.3015	143	0.78221	Marinilabilia
OTU437	0.3067	144	0.79018	Faecalibacterium
OTU453	0.3096	145	0.79215	Paludibacter
OTU309	0.3132	146	0.79587	Sporobacter
OTU380	0.321	147	0.81014	Pseudomonas
OTU367	0.3238	148	0.81169	Faecalibacterium
OTU133	0.3241	149	0.80699	Prevotella
OTU225	0.3246	150	0.80284	Vitellibacter
OTU347	0.3294	151	0.80932	Propionibacterineae
OTU87	0.3324	152	0.81132	Coprococcus
OTU350	0.3391	153	0.82226	Streptococcus
OTU66	0.3455	154	0.83234	Pelomonas
OTU327	0.3464	155	0.82913	Exiguobacterium
OTU364	0.3494	156	0.83094	Lachnospiraceae Incertae Sedis
OTU127	0.3529	157	0.83392	Finegoldia
OTU21	0.3576	158	0.83968	Rikenella
OTU226	0.3623	159	0.84537	Ruminococcaceae Incertae Sedis
OTU150	0.3626	160	0.84078	Lachnospiraceae Incertae Sedis
OTU71	0.3626	161	0.83556	Bacteroides
OTU183	0.364	162	0.8336	Corynebacterineae
OTU445	0.3681	163	0.83782	Lactococcus
OTU213	0.369	164	0.83475	Anaerotruncus
OTU231	0.3705	165	0.83306	Lachnobacterium
OTU119	0.3712	166	0.82961	Lachnospiraceae Incertae Sedis
OTU460	0.3766	167	0.83664	Chryseobacterium

OTU241	0.3767	168	0.83188	Sphingomonas
OTU412	0.3778	169	0.82937	Carnobacteriaceae 1
OTU344	0.3792	170	0.82755	Vibrio
OTU146	0.3819	171	0.82857	Megamonas
OTU114	0.3867	172	0.8341	Micrococcineae
OTU393	0.3888	173	0.83378	Lachnobacterium
OTU417	0.3916	174	0.83496	Lachnospiraceae Incertae Sedis
OTU131	0.3917	175	0.8304	Saprospira
OTU352	0.3921	176	0.82653	Roseburia
OTU358	0.3996	177	0.83758	Lachnospiraceae Incertae Sedis
OTU227	0.4027	178	0.83934	Succinivibrio
OTU53	0.4074	179	0.84439	Bacteroides
OTU36	0.4117	180	0.84856	Coriobacterineae
OTU39	0.4129	181	0.84633	Pseudomonas
OTU97	0.4193	182	0.85473	Bacteroides
OTU89	0.4203	183	0.85208	Faecalibacterium
OTU186	0.4216	184	0.85007	Streptococcus
OTU88	0.4223	185	0.84688	Anaerophaga
OTU283	0.4327	186	0.86307	Lachnospiraceae Incertae Sedis
OTU16	0.4394	187	0.87175	Faecalibacterium
OTU324	0.44	188	0.8683	Coprobacillus
OTU212	0.4402	189	0.8641	Succinivibrio
OTU361	0.4418	190	0.86267	Butyrivibrio
OTU177	0.4429	191	0.86029	Roseburia
OTU379	0.4443	192	0.85852	Lachnospiraceae Incertae Sedis
OTU3	0.4476	193	0.86041	Agrobacterium
OTU319	0.4476	194	0.85598	Coriobacterineae
OTU229	0.4528	195	0.86148	Lachnospiraceae Incertae Sedis
OTU202	0.4564	196	0.8639	Lachnospiraceae Incertae Sedis
OTU311	0.461	197	0.86818	Sphingomonas
OTU265	0.4622	198	0.86604	Aquiflexum
OTU391	0.4654	199	0.86766	Peptostreptococcaceae Incertae Sedis
OTU397	0.4706	200	0.87296	Prevotella
OTU222	0.4779	201	0.88209	Lachnospiraceae Incertae Sedis

OTU40	0.4816	202	0.88452	Bacteroides
OTU196	0.4846	203	0.88565	Lachnospiraceae Incertae Sedis
OTU24	0.4884	204	0.88822	Bryantella
OTU408	0.4951	205	0.89601	Roseburia
OTU153	0.4971	206	0.89526	Fusobacterium
OTU86	0.5011	207	0.89811	Lachnospiraceae Incertae Sedis
OTU326	0.5018	208	0.89504	Clostridiaceae 1
OTU491	0.5047	209	0.8959	Bacteroides
OTU171	0.5061	210	0.89411	Citrobacter
OTU334	0.5071	211	0.89163	Alistipes
OTU194	0.508	212	0.889	Aeromonas
OTU126	0.5122	213	0.89214	Prevotella
OTU237	0.5138	214	0.89075	Dorea
OTU26	0.5169	215	0.89195	Subdoligranulum
OTU60	0.517	216	0.888	Lachnospiraceae Incertae Sedis
OTU52	0.5335	217	0.91211	Ruminococcus
OTU107	0.5352	218	0.91082	Catonella
OTU519	0.5367	219	0.9092	Faecalibacterium
OTU140	0.5398	220	0.9103	Papillibacter
OTU296	0.5432	221	0.91189	Sutterella
OTU49	0.548	222	0.9158	Lachnobacterium
OTU343	0.5663	223	0.94214	Lactobacillus
OTU124	0.5814	224	0.96294	Ruminococcaceae Incertae Sedis
OTU288	0.5881	225	0.96971	Marinilabilia
OTU157	0.5897	226	0.96805	Megamonas
OTU307	0.5901	227	0.96444	Bacteroides
OTU266	0.5921	228	0.96346	Finegoldia
OTU455	0.5928	229	0.96039	Bacteroides
OTU11	0.5944	230	0.95879	Anaerotruncus
OTU94	0.6022	231	0.96717	Turicibacter
OTU109	0.6054	232	0.96812	Bacteroides
OTU85	0.6056	233	0.96428	Roseburia
OTU115	0.6061	234	0.96095	Butyrivibrio
OTU452	0.6141	235	0.96949	Xylanibacter
OTU247	0.6152	236	0.96712	Faecalibacterium

OTU359	0.6155	237	0.9635	Bacteroides
OTU170	0.6263	238	0.97629	Prevotella
OTU341	0.6266	239	0.97267	Lachnospiraceae Incertae Sedis
OTU392	0.6282	240	0.97109	Faecalibacterium
OTU164	0.6284	241	0.96737	Lachnospiraceae Incertae Sedis
OTU57	0.631	242	0.96736	Lachnospiraceae Incertae Sedis
OTU166	0.6318	243	0.9646	Rikenella
OTU219	0.6379	244	0.96992	Parabacteroides
OTU389	0.6418	245	0.97187	Clostridiaceae 1
OTU135	0.6419	246	0.96807	Haemophilus
OTU149	0.6421	247	0.96445	Alkalilimnicola
OTU409	0.6428	248	0.96161	Lachnospiraceae Incertae Sedis
OTU102	0.643	249	0.95804	Peptostreptococcaceae Incertae Sedis
OTU58	0.644	250	0.9557	Burkholderia
OTU118	0.6467	251	0.95588	Parabacteroides
OTU55	0.6552	252	0.9646	Parasporobacterium
OTU328	0.6559	253	0.96181	Lachnospiraceae Incertae Sedis
OTU238	0.6571	254	0.95978	Stenotrophomonas
OTU75	0.6579	255	0.95718	Dorea
OTU429	0.6587	256	0.9546	Peptococcaceae 1
OTU422	0.6675	257	0.96359	Prevotella
OTU122	0.6782	258	0.97524	Rheinheimera
OTU203	0.6874	259	0.98465	Stenotrophomonas
OTU418	0.6879	260	0.98158	Lachnospiraceae Incertae Sedis
OTU463	0.6882	261	0.97825	Prevotella
OTU217	0.6897	262	0.97664	Ruminococcaceae Incertae Sedis
OTU179	0.69	263	0.97335	Dorea
OTU353	0.6943	264	0.9757	Lachnospiraceae Incertae Sedis
OTU20	0.6949	265	0.97286	Lachnospiraceae Incertae Sedis
OTU6	0.6964	266	0.97129	Anaerovorax
OTU456	0.6974	267	0.96905	Bacteroides
OTU158	0.6984	268	0.96681	Alistipes
OTU292	0.6998	269	0.96515	Lachnospiraceae Incertae Sedis
OTU65	0.7036	270	0.9668	Butyrivibrio

OTU345	0.7042	271	0.96405	Lachnospiraceae Incertae Sedis
OTU69	0.7079	272	0.96555	Parabacteroides
OTU267	0.7093	273	0.96392	Sphingobium
OTU474	0.7138	274	0.9665	Lachnospiraceae Incertae Sedis
OTU184	0.7144	275	0.96379	Syntrophococcus
OTU506	0.7161	276	0.96258	Lachnospiraceae Incertae Sedis
OTU44	0.7174	277	0.96085	Roseburia
OTU15	0.7254	278	0.96807	Bacteroides
OTU105	0.7299	279	0.97058	Lachnospiraceae Incertae Sedis
OTU374	0.7312	280	0.96884	Butyrivibrio
OTU285	0.7314	281	0.96566	Methylobacterium
OTU376	0.732	282	0.96302	Anaerotruncus
OTU256	0.7326	283	0.9604	Lachnospiraceae Incertae Sedis
OTU27	0.7346	284	0.95964	Parasporobacterium
OTU423	0.7388	285	0.96174	Anaerovorax
OTU287	0.7472	286	0.96927	Paludibacter
OTU502	0.7498	287	0.96925	Lachnospiraceae Incertae Sedis
OTU274	0.7517	288	0.96834	Lachnospiraceae Incertae Sedis
OTU293	0.7548	289	0.96896	Pseudoalteromonas
OTU101	0.7558	290	0.9669	Faecalibacterium
OTU141	0.761	291	0.97021	Roseburia
OTU129	0.7628	292	0.96917	Comamonas
OTU264	0.7667	293	0.9708	Coprococcus
OTU77	0.7678	294	0.96889	Lachnospiraceae Incertae Sedis
OTU182	0.7731	295	0.97227	Corynebacterineae
OTU355	0.7757	296	0.97225	Lachnospiraceae Incertae Sedis
OTU90	0.777	297	0.9706	Lachnospiraceae Incertae Sedis
OTU29	0.7788	298	0.96958	Lachnospiraceae Incertae Sedis
OTU178	0.7861	299	0.97539	Veillonella
OTU162	0.7889	300	0.97561	Dorea
OTU83	0.7948	301	0.97964	Parabacteroides
OTU25	0.7955	302	0.97725	Acetanaerobacterium
OTU199	0.7962	303	0.97489	Dialister
OTU204	0.808	304	0.98608	Anaerotruncus
OTU354	0.8095	305	0.98467	Lachnospiraceae Incertae Sedis

OTU143	0.8198	306	0.99394	Roseburia
OTU458	0.8218	307	0.99312	Erysipelotrichaceae Incertae Sedis
OTU187	0.8256	308	0.99447	Lachnospiraceae Incertae Sedis
OTU54	0.8309	309	0.99762	Hallella
OTU286	0.8311	310	0.99464	Comamonas
OTU371	0.8371	311	0.9986	Lachnospiraceae Incertae Sedis
OTU4	0.8391	312	0.99778	Micrococcineae
OTU120	0.8398	313	0.99542	Alistipes
OTU401	0.8408	314	0.99343	Peptostreptococcaceae Incertae Sedis
OTU111	0.8414	315	0.99098	Sutterella
OTU50	0.8421	316	0.98867	Pseudomonas
OTU38	0.8472	317	0.99152	Micrococcineae
OTU338	0.8506	318	0.99237	Lachnospiraceae Incertae Sedis
OTU80	0.8517	319	0.99054	Erysipelotrichaceae Incertae Sedis
OTU260	0.8519	320	0.98767	Erysipelotrichaceae Incertae Sedis
OTU32	0.8541	321	0.98714	Lachnobacterium
OTU76	0.8553	322	0.98545	Delftia
OTU56	0.8691	323	0.99825	Enterobacter
OTU313	0.8702	324	0.99643	Faecalibacterium
OTU411	0.871	325	0.99428	Succinispira
OTU47	0.8731	326	0.99362	Azonexus
OTU139	0.8742	327	0.99183	Roseburia
OTU103	0.8747	328	0.98937	Lachnospiraceae Incertae Sedis
OTU198	0.8811	329	0.99358	Sphingobium
OTU70	0.8829	330	0.99259	Faecalibacterium
OTU303	0.8873	331	0.99453	Novosphingobium
OTU356	0.8948	332	0.99991	Turicibacter
OTU407	0.8955	333	0.99769	Parabacteroides
OTU132	0.8999	334	0.99959	Lachnospiraceae Incertae Sedis
OTU79	0.9073	335	1.0048	Subdoligranulum
OTU413	0.9088	336	1.00347	Sporobacter
OTU272	0.9089	337	1.0006	Subdoligranulum
OTU547	0.9101	338	0.99896	Erwinia

OTU176	0.9119	339	0.99798	Coriobacterineae
OTU330	0.913	340	0.99624	Faecalibacterium
OTU363	0.9162	341	0.9968	Coprococcus
OTU396	0.9174	342	0.99519	Anaerotruncus
OTU173	0.9183	343	0.99326	Staphylococcus
OTU268	0.9239	344	0.99642	Lachnospiraceae Incertae Sedis
OTU108	0.926	345	0.99579	Escherichia
OTU17	0.9269	346	0.99387	Bacteroides
OTU9	0.9287	347	0.99293	Erysipelotrichaceae Incertae Sedis
OTU14	0.9289	348	0.99029	Lachnospiraceae Incertae Sedis
OTU148	0.9313	349	0.99001	Roseburia
OTU155	0.9313	350	0.98718	Butyrivibrio
OTU269	0.9376	351	0.99102	Coprococcus
OTU31	0.9397	352	0.99042	Lachnospiraceae Incertae Sedis
OTU499	0.9451	353	0.99329	Lachnospiraceae Incertae Sedis
OTU33	0.9497	354	0.99531	Roseburia
OTU322	0.9515	355	0.99438	Desulfovibrio
OTU235	0.9547	356	0.99493	Sphingomonas
OTU216	0.9582	357	0.99578	Naxibacter
OTU117	0.9598	358	0.99465	Faecalibacterium
OTU2	0.9697	359	1.00211	Faecalibacterium
OTU152	0.9698	360	0.99943	Lachnospiraceae Incertae Sedis
OTU43	0.9713	361	0.99821	Succinispira
OTU270	0.9719	362	0.99606	Bacteroides
OTU181	0.9731	363	0.99455	Ruminococcaceae Incertae Sedis
OTU134	0.9734	364	0.99212	Faecalibacterium
OTU159	0.9739	365	0.98991	Dorea
OTU144	0.9784	366	0.99177	Bacillaceae 1
OTU297	0.9809	367	0.99159	Methylobacterium
OTU403	0.9815	368	0.9895	Coriobacterineae
OTU242	0.9892	369	0.99456	Roseburia
OTU442	0.9918	370	0.99448	Bryantella
OTU400	0.9995	371	0.9995	Simkania

Supplementary Table 6: Regressions on log-normalized abundances of OTUs (97%) vs BMIs of all samples with RDP classifications of consensus sequences at genus level shown. Only OTUs which have at least 1 sequence assigned to them in 25% of the samples are shown. Regression p-Values were corrected for multiple testing[1] using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted rank of the taxon.

OTUname	R^2	p-value	Rank	n*p/R	RDP Genus level Assignment
OTU16	0.12079	0.00320	1	1.18672	Lachnospiraceae Incertae Sedis
OTU492	0.08200	0.01624	2	3.01333	Coriobacterineae
OTU39	0.07881	0.01857	3	2.29692	Coriobacterineae
OTU306	0.07825	0.01901	4	1.76333	Oligotropha
OTU40	0.07472	0.02204	5	1.63559	Lachnospiraceae Incertae Sedis
OTU43	0.07415	0.02257	6	1.39583	Lachnospiraceae Incertae Sedis
OTU305	0.07331	0.02339	7	1.23956	Lachnospiraceae Incertae Sedis
OTU357	0.07070	0.02609	8	1.20976	Coprococcus
OTU4	0.06895	0.02808	9	1.15764	Lachnospiraceae Incertae Sedis
OTU138	0.06863	0.02846	10	1.05595	Simkania
OTU277	0.06168	0.03817	11	1.28733	Lachnospiraceae Incertae Sedis
OTU237	0.05815	0.04432	12	1.37034	Prevotella
OTU131	0.05790	0.04479	13	1.27825	Lachnospiraceae Incertae Sedis
OTU372	0.05470	0.05141	14	1.36242	Allomonas
OTU329	0.05378	0.05339	15	1.32046	Methanohalobium
OTU105	0.05349	0.05406	16	1.25351	Bacteroides
OTU172	0.05309	0.05498	17	1.19992	Marinilabilia
OTU370	0.05290	0.05540	18	1.14185	Lactobacillus
OTU397	0.05190	0.05789	19	1.13039	Peptostreptococcaceae Incertae Sedis
OTU27	0.05132	0.05932	20	1.10034	Lachnospiraceae Incertae Sedis
OTU67	0.05116	0.05973	21	1.05515	Lactobacillus
OTU439	0.05040	0.06178	22	1.0418	Algibacter
OTU110	0.04969	0.06362	23	1.02621	Lachnospiraceae Incertae Sedis
OTU210	0.04921	0.06494	24	1.00386	Allobaculum
OTU380	0.04900	0.06547	25	0.9715	Sporobacter
OTU401	0.04780	0.06903	26	0.98507	Alistipes
OTU204	0.04685	0.07191	27	0.98812	Dialister
OTU288	0.04564	0.07576	28	1.00382	Ruminococcaceae Incertae Sedis
OTU66	0.04482	0.07851	29	1.00441	Streptococcus
OTU432	0.04450	0.07967	30	0.98528	Paludibacter
OTU72	0.04432	0.08022	31	0.96009	Aquabacterium

OTU151	0.04226	0.08778	32	1.01767	Subdoligranulum
OTU167	0.04143	0.09100	33	1.02308	Allobaculum
OTU80	0.04059	0.09443	34	1.03038	Lachnospiraceae Incertae Sedis
OTU153	0.04043	0.09509	35	1.00798	Roseburia
OTU146	0.03945	0.09927	36	1.02302	Vibrio
OTU95	0.03897	0.10141	37	1.01683	Ruminococcus
OTU420	0.03810	0.10547	38	1.02974	Dorea
OTU547	0.03780	0.10677	39	1.01571	Subdoligranulum
OTU352	0.03760	0.10776	40	0.99945	Saprospira
OTU164	0.03704	0.11044	41	0.99931	Faecalibacterium
OTU26	0.03681	0.11160	42	0.98578	Dorea
OTU180	0.03632	0.11402	43	0.98373	Roseburia
OTU373	0.03570	0.11708	44	0.98718	Sporobacter
OTU23	0.03559	0.11780	45	0.97118	Lachnospiraceae Incertae Sedis
OTU230	0.03428	0.12490	46	1.00738	Butyrivibrio
OTU350	0.03420	0.12520	47	0.98831	Coprococcus
OTU88	0.03418	0.12545	48	0.96966	Streptococcus
OTU241	0.03414	0.12570	49	0.95172	Chryseobacterium
OTU309	0.03300	0.13230	50	0.98164	Paludibacter
OTU154	0.03088	0.14566	51	1.05962	Faecalibacterium
OTU499	0.03070	0.14702	52	1.04891	Lachnospiraceae Incertae Sedis
OTU21	0.03053	0.14799	53	1.03595	Finegoldia
OTU452	0.03010	0.15062	54	1.03479	Butyrivibrio
OTU399	0.02990	0.15230	55	1.02734	Ralstonia
OTU96	0.02898	0.15887	56	1.05251	Diaphorobacter
OTU195	0.02838	0.16331	57	1.06294	Pseudoalteromonas
OTU186	0.02821	0.16461	58	1.05293	Faecalibacterium
OTU470	0.02760	0.16933	59	1.06475	Lachnospiraceae Incertae Sedis
OTU84	0.02759	0.16939	60	1.04742	Marinomonas
OTU229	0.02747	0.17030	61	1.03575	Coriobacterineae
OTU566	0.02738	0.17105	62	1.02355	Dorea
OTU98	0.02716	0.17278	63	1.01746	Lachnospiraceae Incertae Sedis
OTU104	0.02705	0.17369	64	1.00683	Syntrophococcus
OTU111	0.02684	0.17532	65	1.00067	Peptostreptococcaceae Incertae Sedis
OTU59	0.02682	0.17553	66	0.98668	Acinetobacter
OTU267	0.02664	0.17697	67	0.97997	Parabacteroides
OTU157	0.02651	0.17809	68	0.97165	Marinilabilia
OTU182	0.02499	0.19123	69	1.02819	Lachnospiraceae Incertae Sedis

OTU231	0.02456	0.19512	70	1.03411	Anaerotruncus
OTU30	0.02451	0.19561	71	1.02215	Bryantella
OTU214	0.02440	0.19663	72	1.0132	Roseburia
OTU538	0.02330	0.20675	73	1.05076	Lachnospiraceae Incertae Sedis
OTU464	0.02320	0.20799	74	1.04277	Marinilabilia
OTU356	0.02290	0.21102	75	1.04383	Novosphingobium
OTU376	0.02220	0.21838	76	1.06602	Methylobacterium
OTU3	0.02217	0.21861	77	1.05332	Lachnospiraceae Incertae Sedis
OTU416	0.02120	0.22887	78	1.0886	Lachnospiraceae Incertae Sedis
OTU358	0.02080	0.23330	79	1.09562	Roseburia
OTU197	0.02052	0.23674	80	1.0979	Lactobacillus
OTU200	0.02050	0.23707	81	1.08584	Helicobacter
OTU495	0.02040	0.23841	82	1.07867	Streptococcus
OTU65	0.01999	0.24295	83	1.08596	Lachnospiraceae Incertae Sedis
OTU454	0.02000	0.24329	84	1.07452	Paludibacter
OTU425	0.01990	0.24367	85	1.06355	Enhydrobacter
OTU46	0.01953	0.24861	86	1.07251	Bacillaceae 1
OTU155	0.01951	0.24887	87	1.06126	Roseburia
OTU240	0.01947	0.24930	88	1.05105	Weissella
OTU266	0.01923	0.25225	89	1.05153	Bacteroides
OTU463	0.01920	0.25304	90	1.04308	Lachnospiraceae Incertae Sedis
OTU107	0.01902	0.25492	91	1.03928	Ruminococcus
OTU101	0.01890	0.25641	92	1.03401	Pseudoalteromonas
OTU102	0.01859	0.26038	93	1.03872	Lachnospiraceae Incertae Sedis
OTU82	0.01851	0.26140	94	1.03169	Roseburia
OTU115	0.01843	0.26242	95	1.02482	Roseburia
OTU51	0.01794	0.26901	96	1.0396	Klebsiella
OTU392	0.01770	0.27267	97	1.04288	Lachnospiraceae Incertae Sedis
OTU198	0.01753	0.27460	98	1.03955	Lachnospiraceae Incertae Sedis
OTU334	0.01747	0.27545	99	1.03225	Citrobacter
OTU423	0.01720	0.27857	100	1.03349	Parasporobacterium
OTU371	0.01710	0.28002	101	1.02858	Comamonas
OTU365	0.01710	0.28007	102	1.01868	Succinispira
OTU367	0.01670	0.28614	103	1.03066	Pseudomonas
OTU378	0.01660	0.28836	104	1.02867	Bacillaceae 1
OTU12	0.01642	0.29042	105	1.02615	Bryantella
OTU47	0.01639	0.29086	106	1.01801	Succinispira
OTU124	0.01633	0.29173	107	1.01152	Lactobacillus

OTU212	0.01631	0.29201	108	1.00313	Coprobacillus
OTU203	0.01613	0.29472	109	1.00314	Rheinheimera
OTU456	0.01590	0.29808	110	1.00533	Anaerovorax
OTU19	0.01563	0.30240	111	1.01072	Syntrophococcus
OTU268	0.01537	0.30653	112	1.01537	Staphylococcus
OTU60	0.01513	0.31036	113	1.01896	Subdoligranulum
OTU50	0.01506	0.31153	114	1.01382	Sutterella
OTU75	0.01487	0.31460	115	1.01494	Stenotrophomonas
OTU192	0.01447	0.32129	116	1.02757	Sphingomonas
OTU36	0.01438	0.32279	117	1.02354	Bacteroides
OTU389	0.01430	0.32348	118	1.01705	Parabacteroides
OTU28	0.01423	0.32534	119	1.01429	Bacteroides
OTU6	0.01415	0.32671	120	1.01009	Lachnospiraceae Incertae Sedis
OTU292	0.01378	0.33313	121	1.0214	Alistipes
OTU282	0.01372	0.33422	122	1.01634	Streptococcus
OTU194	0.01359	0.33650	123	1.01497	Alistipes
OTU15	0.01342	0.33965	124	1.01622	Roseburia
OTU37	0.01340	0.33987	125	1.00874	Cloacibacterium
OTU300	0.01337	0.34042	126	1.00234	Lachnospiraceae Incertae Sedis
OTU165	0.01333	0.34119	127	0.9967	Alistipes
OTU188	0.01329	0.34201	128	0.99129	Lachnospiraceae Incertae Sedis
OTU156	0.01310	0.34551	129	0.99369	Lachnospiraceae Incertae Sedis
OTU304	0.01300	0.34727	130	0.99105	Faecalibacterium
OTU299	0.01299	0.34741	131	0.98388	Lachnospiraceae Incertae Sedis
OTU406	0.01300	0.34761	132	0.97701	Bacteroides
OTU177	0.01289	0.34929	133	0.97433	Butyrivibrio
OTU553	0.01251	0.35656	134	0.98718	Syntrophococcus
OTU190	0.01250	0.35680	135	0.98053	Ruminococcaceae Incertae Sedis
OTU429	0.01210	0.36396	136	0.99285	Dorea
OTU149	0.01212	0.36424	137	0.98637	Haemophilus
OTU24	0.01209	0.36477	138	0.98066	Lachnospiraceae Incertae Sedis
OTU42	0.01196	0.36740	139	0.98061	Prevotella
OTU136	0.01194	0.36780	140	0.97468	Micrococcineae
OTU286	0.01183	0.37015	141	0.97395	Hallella
OTU33	0.01131	0.38093	142	0.99523	Lachnospiraceae Incertae Sedis
OTU455	0.01130	0.38152	143	0.98982	Finegoldia
OTU418	0.01100	0.38698	144	0.997	Stenotrophomonas
OTU91	0.01089	0.38984	145	0.99745	Lactobacillus

OTU256	0.01057	0.39700	146	1.00883	Anaerotruncus
OTU41	0.01030	0.40320	147	1.0176	Subdoligranulum
OTU126	0.01009	0.40791	148	1.02254	Aeromonas
OTU134	0.01007	0.40846	149	1.01703	Ruminococcaceae Incertae Sedis
OTU396	0.00984	0.41387	150	1.02364	Coprococcus
OTU244	0.00967	0.41805	151	1.02712	Prevotella
OTU403	0.00966	0.41823	152	1.02081	Methylobacterium
OTU344	0.00957	0.42046	153	1.01954	Carnobacteriaceae 1
OTU17	0.00947	0.42293	154	1.01888	Escherichia
OTU491	0.00942	0.42407	155	1.01503	Clostridiaceae 1
OTU44	0.00929	0.42739	156	1.01641	Lachnospiraceae Incertae Sedis
OTU29	0.00920	0.42964	157	1.01526	Lachnospiraceae Incertae Sedis
OTU79	0.00897	0.43556	158	1.02274	Lachnospiraceae Incertae Sedis
OTU284	0.00891	0.43701	159	1.01969	Rubritepida
OTU324	0.00890	0.43714	160	1.01362	Faecalibacterium
OTU366	0.00888	0.43768	161	1.00857	Coprococcus
OTU248	0.00884	0.43878	162	1.00486	Lachnospiraceae Incertae Sedis
OTU476	0.00881	0.43963	163	1.00062	Streptococcus
OTU94	0.00876	0.44084	164	0.99725	Anaerotruncus
OTU319	0.00861	0.44499	165	1.00054	Agrobacterium
OTU87	0.00860	0.44510	166	0.99478	Propionibacterineae
OTU11	0.00856	0.44623	167	0.99133	Bacteroides
OTU404	0.00834	0.45223	168	0.99867	Hallella
OTU45	0.00830	0.45326	169	0.99502	Xenohaliotis
OTU61	0.00826	0.45441	170	0.99169	Papillibacter
OTU283	0.00824	0.45488	171	0.9869	Anaerophaga
OTU22	0.00814	0.45764	172	0.98711	Acidovorax
OTU144	0.00814	0.45765	173	0.98144	Dorea
OTU347	0.00805	0.46007	174	0.98094	Vitellibacter
OTU285	0.00766	0.47129	175	0.99914	Butyrivibrio
OTU424	0.00762	0.47244	176	0.99589	Streptococcus
OTU189	0.00739	0.47908	177	1.00417	Acidovorax
OTU417	0.00736	0.47998	178	1.0004	Lachnobacterium
OTU34	0.00734	0.48061	179	0.99612	Dorea
OTU525	0.00724	0.48367	180	0.99691	Catonella
OTU7	0.00717	0.48574	181	0.99564	Bacteroides
OTU32	0.00699	0.49123	182	1.00136	Erysipelotrichaceae Incertae Sedis
OTU168	0.00696	0.49246	183	0.99838	Roseburia

OTU265	0.00694	0.49309	184	0.99422	Sphingomonas
OTU445	0.00686	0.49542	185	0.99352	Corynebacterineae
OTU272	0.00661	0.50356	186	1.00441	Sporobacter
OTU143	0.00640	0.51031	187	1.01243	Lachnospiraceae Incertae Sedis
OTU31	0.00633	0.51268	188	1.01172	Coprococcus
OTU48	0.00615	0.51875	189	1.01829	Bacteroides
OTU184	0.00604	0.52262	190	1.02049	Lachnospiraceae Incertae Sedis
OTU361	0.00599	0.52411	191	1.01804	Succinivibrio
OTU243	0.00590	0.52745	192	1.01919	Anaerotruncus
OTU159	0.00582	0.53006	193	1.01892	Faecalibacterium
OTU400	0.00581	0.53056	194	1.01464	Bryantella
OTU458	0.00574	0.53301	195	1.01409	Roseburia
OTU253	0.00565	0.53639	196	1.01531	Uruburuella
OTU74	0.00557	0.53901	197	1.01509	Ruminococcus
OTU139	0.00546	0.54311	198	1.01765	Azonexus
OTU199	0.00544	0.54396	199	1.01411	Acetanaerobacterium
OTU364	0.00541	0.54523	200	1.0114	Exiguobacterium
OTU129	0.00538	0.54619	201	1.00815	Roseburia
OTU71	0.00534	0.54778	202	1.00608	Lachnospiraceae Incertae Sedis
OTU317	0.00530	0.54939	203	1.00405	Prevotella
OTU52	0.00529	0.54965	204	0.99961	Lachnospiraceae Incertae Sedis
OTU53	0.00528	0.54981	205	0.99502	Succinivibrio
OTU62	0.00497	0.56195	206	1.01205	Ruminococcus
OTU9	0.00494	0.56331	207	1.00961	Bacteroides
OTU311	0.00484	0.56729	208	1.01184	Lachnospiraceae Incertae Sedis
OTU76	0.00483	0.56755	209	1.00747	Lachnobacterium
OTU89	0.00483	0.56764	210	1.00282	Bacteroides
OTU216	0.00471	0.57232	211	1.0063	Sphingomonas
OTU58	0.00470	0.57286	212	1.00251	Peptostreptococcaceae Incertae Sedis
OTU133	0.00469	0.57321	213	0.99841	Faecalibacterium
OTU493	0.00435	0.58737	214	1.01829	Lachnospiraceae Incertae Sedis
OTU327	0.00434	0.58810	215	1.01482	Pelomonas
OTU49	0.00427	0.59075	216	1.01466	Sutterella
OTU242	0.00427	0.59078	217	1.01005	Coriobacterineae
OTU359	0.00427	0.59097	218	1.00573	Faecalibacterium
OTU316	0.00424	0.59231	219	1.00341	Alistipes
OTU73	0.00421	0.59368	220	1.00116	Lactococcus
OTU2	0.00416	0.59600	221	1.00053	Faecalibacterium

OTU484	0.00410	0.59856	222	1.00029	Effluviibacter
OTU297	0.00408	0.59957	223	0.99749	Bacillaceae 1
OTU150	0.00406	0.60032	224	0.99428	Ruminococcaceae Incertae Sedis
OTU239	0.00388	0.60851	225	1.00337	Succinispira
OTU205	0.00376	0.61391	226	1.00778	Erysipelotrichaceae Incertae Sedis
OTU38	0.00375	0.61436	227	1.00408	Pseudomonas
OTU117	0.00370	0.61669	228	1.00347	Naxibacter
OTU274	0.00366	0.61881	229	1.00253	Lachnospiraceae Incertae Sedis
OTU341	0.00361	0.62128	230	1.00214	Prevotella
OTU170	0.00359	0.62208	231	0.9991	Bacteroides
OTU207	0.00358	0.62246	232	0.9954	Succinispira
OTU90	0.00346	0.62846	233	1.00069	Lachnospiraceae Incertae Sedis
OTU296	0.00337	0.63322	234	1.00396	Papillibacter
OTU238	0.00333	0.63519	235	1.00279	Lachnospiraceae Incertae Sedis
OTU227	0.00333	0.63529	236	0.9987	Lachnospiraceae Incertae Sedis
OTU374	0.00321	0.64151	237	1.00423	Lachnospiraceae Incertae Sedis
OTU114	0.00320	0.64157	238	1.0001	Megamonas
OTU152	0.00316	0.64412	239	0.99986	Faecalibacterium
OTU395	0.00315	0.64466	240	0.99653	Subdoligranulum
OTU326	0.00296	0.65473	241	1.0079	Lachnospiraceae Incertae Sedis
OTU226	0.00293	0.65630	242	1.00615	Rikenella
OTU56	0.00271	0.66884	243	1.02115	Delftia
OTU57	0.00270	0.66907	244	1.01731	Lachnospiraceae Incertae Sedis
OTU249	0.00269	0.66999	245	1.01456	Faecalibacterium
OTU187	0.00262	0.67379	246	1.01616	Erysipelotrichaceae Incertae Sedis
OTU173	0.00255	0.67803	247	1.01842	Anaerotruncus
OTU77	0.00255	0.67813	248	1.01446	Coprococcus
OTU519	0.00254	0.67847	249	1.01089	Catonella
OTU313	0.00252	0.67991	250	1.00899	Enterobacter
OTU233	0.00249	0.68143	251	1.00722	Syntrophococcus
OTU179	0.00241	0.68654	252	1.01074	Ruminococcaceae Incertae Sedis
OTU506	0.00237	0.68930	253	1.01079	Syntrophococcus
OTU103	0.00225	0.69653	254	1.01738	Roseburia
OTU407	0.00223	0.69779	255	1.01521	Turicibacter
OTU269	0.00222	0.69851	256	1.0123	Butyrivibrio
OTU222	0.00220	0.69989	257	1.01035	Prevotella
OTU193	0.00215	0.70341	258	1.01149	Xylanibacter

OTU132	0.00199	0.71391	259	1.02263	Parabacteroides
OTU411	0.00192	0.71867	260	1.02548	Faecalibacterium
OTU109	0.00191	0.71934	261	1.02251	Turicibacter
OTU181	0.00189	0.72104	262	1.02101	Bacteroides
OTU413	0.00183	0.72484	263	1.0225	Subdoligranulum
OTU508	0.00183	0.72503	264	1.01889	Lachnospiraceae Incertae Sedis
OTU127	0.00172	0.73283	265	1.02596	Lachnospiraceae Incertae Sedis
OTU219	0.00164	0.73945	266	y.03134	Rikenella
OTU202	0.00152	0.74899	267	1.04073	Lachnospiraceae Incertae Sedis
OTU158	0.00145	0.75455	268	1.04455	Bacteroides
OTU113	0.00145	0.75468	269	1.04084	Rikenella
OTU291	0.00143	0.75607	270	1.0389	Syntrophococcus
OTU35	0.00138	0.75983	271	1.0402	Bryantella
OTU69	0.00138	0.76032	272	1.03706	Lachnospiraceae Incertae Sedis
OTU360	0.00138	0.76046	273	1.03345	Faecalibacterium
OTU270	0.00137	0.76063	274	1.0299	Succinispira
OTU569	0.00136	0.76170	275	1.0276	Erwinia
OTU148	0.00121	0.77482	276	1.04151	Lachnospiraceae Incertae Sedis
OTU206	0.00118	0.77735	277	1.04114	Paludibacter
OTU338	0.00110	0.78478	278	1.04732	Micrococcineae
OTU25	0.00110	0.78564	279	1.04471	Parabacteroides
OTU108	0.00109	0.78588	280	1.04129	Lachnospiraceae Incertae Sedis
OTU328	0.00104	0.79060	281	1.04382	Parasporobacterium
OTU419	0.00104	0.79110	282	1.04078	Micrococcineae
OTU225	0.00104	0.79121	283	1.03725	Prevotella
OTU123	0.00104	0.79133	284	1.03375	Papillibacter
OTU460	0.00098	0.79703	285	1.03754	Lachnospiraceae Incertae Sedis
OTU70	0.00094	0.80105	286	1.03913	Sphingobium
OTU1	0.00093	0.80167	287	1.0363	Bacteroides
OTU387	0.00093	0.80206	288	1.03321	Coprococcus
OTU345	0.00090	0.80526	289	1.03374	Butyrivibrio
OTU137	0.00090	0.80547	290	1.03045	Prevotella
OTU10	0.00089	0.80605	291	1.02764	Coprobacillus
OTU312	0.00083	0.81254	292	1.03237	Coriobacterineae
OTU307	0.00080	0.81611	293	1.03337	Megamonas
OTU353	0.00079	0.81796	294	1.03218	Dorea
OTU196	0.00078	0.81801	295	1.02875	Bacteroides
OTU8	0.00078	0.81824	296	1.02556	Dorea

OTU178	0.00072	0.82507	297	1.03064	Lachnospiraceae Incertae Sedis
OTU106	0.00072	0.82581	298	1.02811	Lachnospiraceae Incertae Sedis
OTU437	0.00071	0.82714	299	1.02632	Marinilabilia
OTU393	0.00069	0.82865	300	1.02476	Micrococcineae
OTU502	0.00067	0.83190	301	1.02536	Paludibacter
OTU349	0.00066	0.83311	302	1.02345	Syntrophococcus
OTU343	0.00065	0.83398	303	1.02115	Lachnobacterium
OTU354	0.00064	0.83515	304	1.01921	Anaerotruncus
OTU120	0.00064	0.83562	305	1.01644	Micrococcineae
OTU368	0.00060	0.83993	306	1.01835	Ruminococcaceae Incertae Sedis
OTU330	0.00060	0.84109	307	1.01643	Coriobacterineae
OTU18	0.00058	0.84311	308	1.01557	Faecalibacterium
OTU379	0.00055	0.84661	309	1.01647	Roseburia
OTU355	0.00052	0.85194	310	1.01958	Corynebacterineae
OTU169	0.00048	0.85685	311	1.02216	Streptococcus
OTU217	0.00044	0.86299	312	1.02619	Prevotella
OTU97	0.00044	0.86362	313	1.02365	Pseudomonas
OTU315	0.00043	0.86508	314	1.02211	Coriobacterineae
OTU453	0.00041	0.86851	315	1.02292	Faecalibacterium
OTU293	0.00041	0.86858	316	1.01975	Lachnospiraceae Incertae Sedis
OTU160	0.00039	0.87159	317	1.02006	Lachnospiraceae Incertae Sedis
OTU93	0.00038	0.87290	318	1.01839	Alistipes
OTU303	0.00037	0.87374	319	1.01617	Faecalibacterium
OTU128	0.00036	0.87555	320	1.01509	Prevotella
OTU86	0.00035	0.87754	321	1.01423	Fusobacterium
OTU264	0.00035	0.87829	322	1.01195	Comamonas
OTU171	0.00034	0.87891	323	1.00952	Bacteroides
OTU100	0.00032	0.88369	324	1.01187	Xylanibacter
OTU176	0.00032	0.88369	325	1.00877	Erwinia
OTU235	0.00030	0.88760	326	1.01013	Desulfovibrio
OTU142	0.00027	0.89298	327	1.01314	Lachnospiraceae Incertae Sedis
OTU183	0.00025	0.89598	328	1.01344	Bacteroides
OTU391	0.00024	0.89806	329	1.01271	Aquiflexum
OTU85	0.00024	0.89815	330	1.00974	Bacteroides
OTU224	0.00023	0.90135	331	1.01028	Prevotella
OTU55	0.00023	0.90176	332	1.00769	Parabacteroides
OTU166	0.00022	0.90242	333	1.00539	Lachnospiraceae Incertae Sedis
OTU322	0.00021	0.90433	334	1.00451	Roseburia

OTU14	0.00020	0.90785	335	1.0054	Erysipelotrichaceae Incertae Sedis
OTU408	0.00019	0.90951	336	1.00425	Bryantella
OTU54	0.00018	0.91151	337	1.00347	Lachnospiraceae Incertae Sedis
OTU64	0.00017	0.91495	338	1.00428	Erwinia
OTU83	0.00017	0.91541	339	1.00182	Dorea
OTU68	0.00016	0.91804	340	1.00175	Dorea
OTU5	0.00015	0.92125	341	1.0023	Sphingomonas
OTU145	0.00014	0.92320	342	1.00149	Afipia
OTU119	0.00014	0.92370	343	0.9991	Lachnobacterium
OTU442	0.00011	0.93035	344	1.00337	Roseburia
OTU412	0.00011	0.93055	345	1.00068	Sphingomonas
OTU474	0.00011	0.93058	346	0.99781	Sphingobium
OTU20	0.00011	0.93225	347	0.99673	Lachnospiraceae Incertae Sedis
OTU254	0.00010	0.93343	348	0.99512	Lachnospiraceae Incertae Sedis
OTU260	0.00010	0.93363	349	0.99248	Erysipelotrichaceae Incertae Sedis
OTU287	0.00010	0.93561	350	0.99175	Anaerovorax
OTU250	0.00009	0.93893	351	0.99243	Paludibacter
OTU422	0.00009	0.93947	352	0.99018	Peptococcaceae 1
OTU140	0.00008	0.94086	353	0.98883	Faecalibacterium
OTU421	0.00008	0.94289	354	0.98817	Streptococcus
OTU161	0.00006	0.94925	355	0.99203	Prevotella
OTU135	0.00006	0.94978	356	0.9898	Clostridiaceae 1
OTU375	0.00005	0.95255	357	0.98991	Pseudomonas
OTU191	0.00005	0.95294	358	0.98754	Subdoligranulum
OTU122	0.00004	0.95860	359	0.99064	Prevotella
OTU162	0.00004	0.95894	360	0.98824	Veillonella
OTU501	0.00004	0.95986	361	0.98645	Ruminococcaceae Incertae Sedis
OTU275	0.00004	0.96063	362	0.98452	Lachnospiraceae Incertae Sedis
OTU213	0.00004	0.96094	363	0.98212	Lactococcus
OTU141	0.00003	0.96233	364	0.98084	Faecalibacterium
OTU363	0.00003	0.96649	365	0.98238	Faecalibacterium
OTU130	0.00002	0.97345	366	0.98675	Lachnospiraceae Incertae Sedis
OTU409	0.00001	0.97609	367	0.98673	Alkalilimnicola
OTU471	0.00001	0.97787	368	0.98584	Lachnospiraceae Incertae Sedis
OTU247	0.00000	0.98560	369	0.99095	Xylanibacter
OTU118	0.00000	0.99027	370	0.99295	Burkholderia
OTU92	0.00000	0.99641	371	0.99641	Rubrobacterineae

Supplementary Table 7: Regressions on log-normalized abundances of OTUs (97%) vs. WHRs of all samples with RDP classification of consensus sequences at genus level shown. Only OTUs which have at least 1 sequence assigned to them in 25% of the samples are shown. Regression p-Values were corrected for multiple testing[1] using (n*p)/R where n = total number of taxa tested, p = raw p-Value and R = sorted rank of the taxon.

OTUname	R^2	p-value	Rank	n*p/R	RDP Genus level Assignment
OTU4	0.16058	0.00053	1	0.19811	Lachnospiraceae Incertae Sedis
OTU492	0.16000	0.00054	2	0.09998	Coriobacterineae
OTU305	0.15413	0.00071	3	0.08756	Lachnospiraceae Incertae Sedis
OTU79	0.09585	0.00861	4	0.79813	Lachnospiraceae Incertae Sedis
OTU476	0.09510	0.00890	5	0.66061	Streptococcus
OTU132	0.09057	0.01076	6	0.66561	Parabacteroides
OTU123	0.09019	0.01094	7	0.57987	Papillibacter
OTU31	0.07537	0.02050	8	0.95086	Coprococcus
OTU249	0.07253	0.02314	9	0.9537	Faecalibacterium
OTU416	0.06910	0.02679	10	0.99377	Lachnospiraceae Incertae Sedis
OTU471	0.06680	0.02958	11	0.99774	Lachnospiraceae Incertae Sedis
OTU3	0.06375	0.03364	12	1.04016	Lachnospiraceae Incertae Sedis
OTU54	0.06336	0.03421	13	0.97625	Lachnospiraceae Incertae Sedis
OTU36	0.06000	0.03952	14	1.0472	Bacteroides
OTU282	0.05870	0.04177	15	1.03316	Streptococcus
OTU162	0.05520	0.04858	16	1.12656	Veillonella
OTU11	0.05483	0.04936	17	1.07724	Bacteroides
OTU420	0.05420	0.05065	18	1.04393	Dorea
OTU2	0.05334	0.05265	19	1.02803	Faecalibacterium
OTU306	0.05307	0.05327	20	0.98819	Oligotropha
OTU14	0.05298	0.05347	21	0.94458	Erysipelotrichaceae Incertae Sedis
OTU122	0.04952	0.06214	22	1.04792	Prevotella
OTU65	0.04587	0.07291	23	1.17604	Lachnospiraceae Incertae Sedis
OTU242	0.04413	0.07870	24	1.21653	Coriobacterineae
OTU199	0.04234	0.08517	25	1.26385	Acetanaerobacterium
OTU330	0.04207	0.08618	26	1.22971	Coriobacterineae
OTU239	0.04187	0.08696	27	1.19491	Succinispira
OTU197	0.04077	0.09130	28	1.20971	Lactobacillus
OTU229	0.03893	0.09909	29	1.26763	Coriobacterineae
OTU149	0.03824	0.10219	30	1.26381	Haemophilus
OTU28	0.03786	0.10396	31	1.24416	Bacteroides
OTU49	0.03752	0.10553	32	1.2235	Sutterella
OTU237	0.03741	0.10605	33	1.19224	Prevotella

OTU29	0.03739	0.10616	34	1.15839	Lachnospiraceae Incertae Sedis
OTU27	0.03664	0.10980	35	1.16391	Lachnospiraceae Incertae Sedis
OTU74	0.03641	0.11095	36	1.14341	Ruminococcus
OTU284	0.03627	0.11165	37	1.11954	Rubritepida
OTU198	0.03622	0.11189	38	1.09235	Lachnospiraceae Incertae Sedis
OTU329	0.03581	0.11399	39	1.08437	Methanohalobium
OTU283	0.03545	0.11583	40	1.07435	Anaerophaga
OTU72	0.03517	0.11730	41	1.06145	Aquabacterium
OTU309	0.03504	0.11804	42	1.04269	Paludibacter
OTU59	0.03413	0.12299	43	1.06115	Acinetobacter
OTU470	0.03410	0.12300	44	1.03708	Lachnospiraceae Incertae Sedis
OTU173	0.03391	0.12420	45	1.02394	Anaerotruncus
OTU454	0.03280	0.13051	46	1.05262	Paludibacter
OTU16	0.03271	0.13118	47	1.03546	Lachnospiraceae Incertae Sedis
OTU356	0.03220	0.13429	48	1.03794	Novosphingobium
OTU46	0.03150	0.13869	49	1.05007	Bacillaceae 1
OTU98	0.03113	0.14105	50	1.04662	Lachnospiraceae Incertae Sedis
OTU288	0.03108	0.14138	51	1.02847	Ruminococcaceae Incertae Sedis
OTU474	0.03040	0.14608	52	1.04224	Sphingobium
OTU104	0.02913	0.15475	53	1.08326	Syntrophococcus
OTU429	0.02890	0.15635	54	1.07418	Dorea
OTU41	0.02856	0.15889	55	1.07178	Subdoligranulum
OTU117	0.02834	0.16052	56	1.06347	Naxibacter
OTU96	0.02828	0.16096	57	1.04767	Diaphorobacter
OTU143	0.02795	0.16346	58	1.04555	Lachnospiraceae Incertae Sedis
OTU367	0.02760	0.16620	59	1.04507	Pseudomonas
OTU34	0.02734	0.16820	60	1.04003	Dorea
OTU200	0.02721	0.16926	61	1.02946	Helicobacter
OTU525	0.02660	0.17395	62	1.04092	Catonella
OTU42	0.02657	0.17443	63	1.02721	Prevotella
OTU376	0.02630	0.17634	64	1.02221	Methylobacterium
OTU128	0.02590	0.18004	65	1.02761	Prevotella
OTU368	0.02540	0.18463	66	1.03784	Ruminococcaceae Incertae Sedis
OTU58	0.02536	0.18466	67	1.0225	Peptostreptococcaceae Incertae Sedis
OTU349	0.02528	0.18537	68	1.01137	Syntrophococcus
OTU268	0.02473	0.19030	69	1.02319	Staphylococcus
OTU88	0.02472	0.19038	70	1.00902	Streptococcus
OTU327	0.02412	0.19593	71	1.02381	Pelomonas
OTU370	0.02370	0.19945	72	1.02772	Lactobacillus
OTU134	0.02349	0.20191	73	1.02617	Ruminococcaceae Incertae Sedis

OTU150	0.02343	0.20256	74	1.01552	Ruminococcaceae Incertae Sedis
OTU203	0.02326	0.20419	75	1.01007	Rheinheimera
OTU391	0.02320	0.20459	76	0.99874	Aquiflexum
OTU363	0.02250	0.21188	77	1.02088	Faecalibacterium
OTU413	0.02250	0.21201	78	1.00838	Subdoligranulum
OTU231	0.02211	0.21589	79	1.01386	Anaerotruncus
OTU66	0.02207	0.21626	80	1.00289	Streptococcus
OTU350	0.02190	0.21793	81	0.99816	Coprococcus
OTU269	0.02141	0.22340	82	1.01077	Butyrivibrio
OTU131	0.02120	0.22564	83	1.0086	Lachnospiraceae Incertae Sedis
OTU61	0.02022	0.23682	84	1.04596	Papillibacter
OTU235	0.02020	0.23709	85	1.03484	Desulfovibrio
OTU343	0.02019	0.23722	86	1.02337	Lachnobacterium
OTU172	0.01971	0.24294	87	1.03601	Marinilabilia
OTU299	0.01952	0.24515	88	1.03353	Lachnospiraceae Incertae Sedis
OTU425	0.01920	0.24895	89	1.03778	Enhydrobacter
OTU213	0.01908	0.25071	90	1.0335	Lactococcus
OTU25	0.01902	0.25143	91	1.02507	Parabacteroides
OTU140	0.01892	0.25267	92	1.01892	Faecalibacterium
OTU403	0.01870	0.25498	93	1.01717	Methylobacterium
OTU204	0.01831	0.26054	94	1.02831	Dialister
OTU157	0.01811	0.26320	95	1.02788	Marinilabilia
OTU359	0.01780	0.26799	96	1.03568	Faecalibacterium
OTU214	0.01759	0.27025	97	1.03365	Roseburia
OTU566	0.01752	0.27111	98	1.02633	Dorea
OTU37	0.01740	0.27290	99	1.02267	Cloacibacterium
OTU371	0.01740	0.27331	100	1.01397	Comamonas
OTU18	0.01721	0.27546	101	1.01184	Faecalibacterium
OTU146	0.01721	0.27553	102	1.00216	Vibrio
OTU354	0.01710	0.27690	103	0.99738	Anaerotruncus
OTU357	0.01690	0.27932	104	0.99642	Coprococcus
OTU334	0.01680	0.28133	105	0.99405	Citrobacter
OTU352	0.01630	0.28894	106	1.0113	Saprospira
OTU274	0.01605	0.29249	107	1.01413	Lachnospiraceae Incertae Sedis
OTU326	0.01598	0.29346	108	1.0081	Lachnospiraceae Incertae Sedis
OTU1	0.01594	0.29407	109	1.00092	Bacteroides
OTU191	0.01560	0.29941	110	1.00983	Subdoligranulum
OTU40	0.01507	0.30780	111	1.02877	Lachnospiraceae Incertae Sedis
OTU226	0.01504	0.30832	112	1.0213	Rikenella
OTU48	0.01480	0.31210	113	1.02469	Bacteroides

OTU39	0.01476	0.31278	114	1.0179	Coriobacterineae
OTU364	0.01470	0.31323	115	1.0105	Exiguobacterium
OTU178	0.01467	0.31438	116	1.00547	Lachnospiraceae Incertae Sedis
OTU113	0.01446	0.31778	117	1.00765	Rikenella
OTU32	0.01434	0.31990	118	1.0058	Erysipelotrichaceae Incertae Sedis
OTU296	0.01416	0.32295	119	1.00685	Papillibacter
OTU153	0.01415	0.32311	120	0.99894	Roseburia
OTU502	0.01410	0.32410	121	0.99373	Paludibacter
OTU324	0.01390	0.32745	122	0.99577	Faecalibacterium
OTU110	0.01387	0.32801	123	0.98936	Lachnospiraceae Incertae Sedis
OTU315	0.01382	0.32888	124	0.98397	Coriobacterineae
OTU102	0.01344	0.33568	125	0.99631	Lachnospiraceae Incertae Sedis
OTU193	0.01339	0.33664	126	0.99121	Xylanibacter
OTU15	0.01337	0.33695	127	0.98432	Roseburia
OTU103	0.01314	0.34116	128	0.98882	Roseburia
OTU184	0.01280	0.34746	129	0.99928	Lachnospiraceae Incertae Sedis
OTU169	0.01267	0.34993	130	0.99865	Streptococcus
OTU23	0.01263	0.35081	131	0.99351	Lachnospiraceae Incertae Sedis
OTU53	0.01249	0.35340	132	0.99326	Succinivibrio
OTU247	0.01237	0.35585	133	0.99263	Xylanibacter
OTU7	0.01232	0.35687	134	0.98806	Bacteroides
OTU20	0.01229	0.35738	135	0.98213	Lachnospiraceae Incertae Sedis
OTU77	0.01223	0.35855	136	0.97811	Coprococcus
OTU358	0.01210	0.36096	137	0.97748	Roseburia
OTU423	0.01200	0.36253	138	0.97464	Parasporobacterium
OTU508	0.01190	0.36508	139	0.97443	Lachnospiraceae Incertae Sedis
OTU322	0.01160	0.37141	140	0.98423	Roseburia
OTU84	0.01152	0.37297	141	0.98135	Marinomonas
OTU210	0.01152	0.37298	142	0.97447	Allobaculum
OTU22	0.01147	0.37410	143	0.97058	Acidovorax
OTU380	0.01120	0.37870	144	0.97568	Sporobacter
OTU553	0.01109	0.38216	145	0.97781	Syntrophococcus
OTU389	0.01090	0.38598	146	0.9808	Parabacteroides
OTU392	0.01060	0.39195	147	0.98921	Lachnospiraceae Incertae Sedis
OTU344	0.01063	0.39231	148	0.98343	Carnobacteriaceae 1
OTU506	0.01060	0.39366	149	0.98018	Syntrophococcus
OTU177	0.01020	0.40194	150	0.99414	Butyrivibrio
OTU399	0.01000	0.40554	151	0.99638	Ralstonia
OTU300	0.00991	0.40888	152	0.99798	Lachnospiraceae Incertae Sedis
OTU316	0.00972	0.41345	153	1.00255	Alistipes

OTU456	0.00959	0.41661	154	1.00364	Anaerovorax
OTU293	0.00946	0.41960	155	1.00433	Lachnospiraceae Incertae Sedis
OTU21	0.00935	0.42250	156	1.0048	Finegoldia
OTU361	0.00922	0.42574	157	1.00604	Succinivibrio
OTU202	0.00914	0.42775	158	1.00439	Lachnospiraceae Incertae Sedis
OTU366	0.00895	0.43267	159	1.00957	Coprococcus
OTU35	0.00884	0.43540	160	1.00958	Bryantella
OTU275	0.00833	0.44901	161	1.03468	Lachnospiraceae Incertae Sedis
OTU126	0.00830	0.44997	162	1.03049	Aeromonas
OTU189	0.00828	0.45054	163	1.02547	Acidovorax
OTU158	0.00826	0.45096	164	1.02016	Bacteroides
OTU43	0.00807	0.45634	165	1.02607	Lachnospiraceae Incertae Sedis
OTU105	0.00801	0.45787	166	1.02332	Bacteroides
OTU9	0.00797	0.45913	167	1.01997	Bacteroides
OTU297	0.00745	0.47430	168	1.04742	Bacillaceae 1
OTU80	0.00741	0.47546	169	1.04376	Lachnospiraceae Incertae Sedis
OTU277	0.00732	0.47801	170	1.04318	Lachnospiraceae Incertae Sedis
OTU395	0.00727	0.47963	171	1.04061	Subdoligranulum
OTU365	0.00727	0.47972	172	1.03475	Succinispira
OTU67	0.00726	0.47982	173	1.02898	Lactobacillus
OTU372	0.00714	0.48370	174	1.03133	Allomonas
OTU419	0.00701	0.48759	175	1.03368	Micrococcineae
OTU101	0.00697	0.48875	176	1.03027	Pseudoalteromonas
OTU10	0.00692	0.49053	177	1.02818	Coprobacillus
OTU154	0.00685	0.49266	178	1.02683	Faecalibacterium
OTU93	0.00677	0.49515	179	1.02626	Alistipes
OTU62	0.00672	0.49684	180	1.02404	Ruminococcus
OTU404	0.00645	0.50544	181	1.03602	Hallella
OTU406	0.00645	0.50564	182	1.03073	Bacteroides
OTU241	0.00635	0.50892	183	1.03175	Chryseobacterium
OTU151	0.00634	0.50932	184	1.02695	Subdoligranulum
OTU307	0.00629	0.51093	185	1.02461	Megamonas
OTU155	0.00621	0.51362	186	1.02448	Roseburia
OTU264	0.00619	0.51413	187	1.02	Comamonas
OTU124	0.00607	0.51856	188	1.02333	Lactobacillus
OTU227	0.00595	0.52273	189	1.0261	Lachnospiraceae Incertae Sedis
OTU12	0.00581	0.52761	190	1.03023	Bryantella
OTU442	0.00580	0.52800	191	1.02559	Roseburia
OTU187	0.00572	0.53082	192	1.0257	Erysipelotrichaceae Incertae Sedis
OTU45	0.00570	0.53138	193	1.02145	Xenohaliotis

OTU240	0.00562	0.53429	194	1.02176	Weissella
OTU95	0.00533	0.54511	195	1.03711	Ruminococcus
OTU87	0.00532	0.54540	196	1.03237	Propionibacterineae
OTU129	0.00524	0.54849	197	1.03295	Roseburia
OTU243	0.00519	0.55054	198	1.03156	Anaerotruncus
OTU133	0.00517	0.55109	199	1.02741	Faecalibacterium
OTU401	0.00516	0.55153	200	1.0231	Alistipes
OTU421	0.00511	0.55354	201	1.02171	Streptococcus
OTU152	0.00508	0.55466	202	1.01871	Faecalibacterium
OTU253	0.00503	0.55669	203	1.0174	Uruburuella
OTU171	0.00501	0.55767	204	1.01419	Bacteroides
OTU109	0.00499	0.55827	205	1.01034	Turicibacter
OTU445	0.00483	0.56471	206	1.01703	Corynebacterineae
OTU137	0.00471	0.56939	207	1.0205	Prevotella
OTU100	0.00458	0.57482	208	1.02527	Xylanibacter
OTU130	0.00454	0.57648	209	1.02333	Lachnospiraceae Incertae Sedis
OTU328	0.00454	0.57671	210	1.01885	Parasporobacterium
OTU378	0.00452	0.57768	211	1.01573	Bacillaceae 1
OTU183	0.00442	0.58181	212	1.01816	Bacteroides
OTU26	0.00438	0.58344	213	1.01623	Dorea
OTU432	0.00438	0.58352	214	1.01161	Paludibacter
OTU317	0.00426	0.58867	215	1.01579	Prevotella
OTU256	0.00424	0.58935	216	1.01226	Anaerotruncus
OTU353	0.00424	0.58952	217	1.00789	Dorea
OTU114	0.00424	0.58957	218	1.00335	Megamonas
OTU453	0.00421	0.59104	219	1.00126	Faecalibacterium
OTU94	0.00411	0.59542	220	1.0041	Anaerotruncus
OTU460	0.00405	0.59791	221	1.00373	Lachnospiraceae Incertae Sedis
OTU194	0.00393	0.60353	222	1.0086	Alistipes
OTU159	0.00384	0.60749	223	1.01066	Faecalibacterium
OTU141	0.00369	0.61465	224	1.01802	Faecalibacterium
OTU90	0.00369	0.61470	225	1.01357	Lachnospiraceae Incertae Sedis
OTU217	0.00367	0.61566	226	1.01066	Prevotella
OTU397	0.00363	0.61788	227	1.00983	Peptostreptococcaceae Incertae Sedis
OTU374	0.00353	0.62263	228	1.01314	Lachnospiraceae Incertae Sedis
OTU148	0.00345	0.62636	229	1.01476	Lachnospiraceae Incertae Sedis
OTU19	0.00337	0.63044	230	1.01693	Syntrophococcus
OTU422	0.00334	0.63195	231	1.01495	Peptococcaceae 1
OTU418	0.00309	0.64516	232	1.0317	Stenotrophomonas
OTU33	0.00308	0.64573	233	1.02818	Lachnospiraceae Incertae Sedis

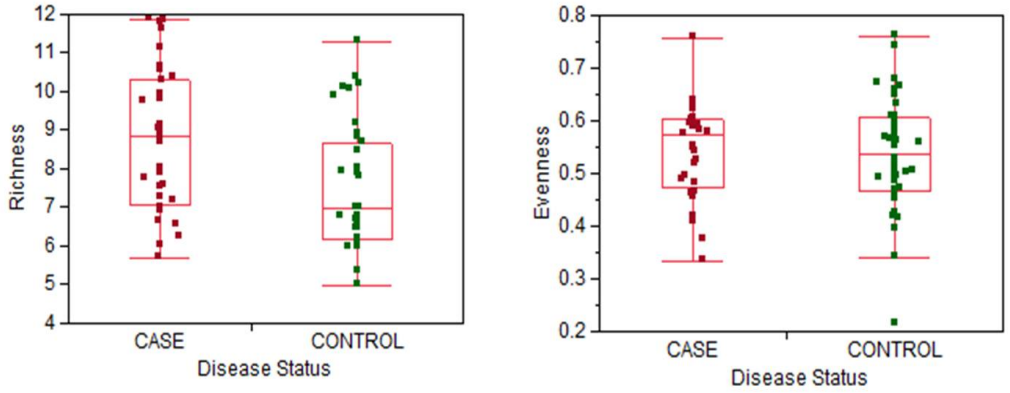
OTU38	0.00307	0.64629	234	1.02467	Pseudomonas
OTU75	0.00303	0.64841	235	1.02366	Stenotrophomonas
OTU138	0.00287	0.65749	236	1.0336	Simkania
OTU396	0.00276	0.66333	237	1.03838	Coprococcus
OTU311	0.00274	0.66482	238	1.03634	Lachnospiraceae Incertae Sedis
OTU73	0.00270	0.66679	239	1.03505	Lactococcus
OTU455	0.00255	0.67552	240	1.04424	Finegoldia
OTU407	0.00250	0.67877	241	1.04492	Turicibacter
OTU238	0.00247	0.68035	242	1.04301	Lachnospiraceae Incertae Sedis
OTU501	0.00245	0.68189	243	1.04107	Ruminococcaceae Incertae Sedis
OTU6	0.00241	0.68430	244	1.04047	Lachnospiraceae Incertae Sedis
OTU225	0.00236	0.68772	245	1.04141	Prevotella
OTU347	0.00233	0.68910	246	1.03925	Vitellibacter
OTU355	0.00229	0.69179	247	1.03909	Corynebacterineae
OTU135	0.00229	0.69192	248	1.03508	Clostridiaceae 1
OTU8	0.00225	0.69454	249	1.03483	Dorea
OTU417	0.00225	0.69474	250	1.031	Lachnobacterium
OTU30	0.00217	0.69963	251	1.03412	Bryantella
OTU484	0.00210	0.70453	252	1.03722	Effluviibacter
OTU265	0.00199	0.71215	253	1.04431	Sphingomonas
OTU24	0.00195	0.71462	254	1.04379	Lachnospiraceae Incertae Sedis
OTU224	0.00194	0.71545	255	1.04092	Prevotella
OTU219	0.00181	0.72457	256	1.05005	Rikenella
OTU499	0.00174	0.72958	257	1.0532	Lachnospiraceae Incertae Sedis
OTU192	0.00171	0.73229	258	1.05302	Sphingomonas
OTU212	0.00169	0.73349	259	1.05067	Coprobacillus
OTU312	0.00164	0.73726	260	1.05202	Coriobacterineae
OTU55	0.00163	0.73794	261	1.04895	Parabacteroides
OTU286	0.00163	0.73815	262	1.04524	Hallella
OTU142	0.00158	0.74217	263	1.04693	Lachnospiraceae Incertae Sedis
OTU106	0.00155	0.74467	264	1.04648	Lachnospiraceae Incertae Sedis
OTU161	0.00144	0.75323	265	1.05452	Prevotella
OTU165	0.00141	0.75569	266	1.05399	Alistipes
OTU186	0.00139	0.75723	267	1.05218	Faecalibacterium
OTU439	0.00136	0.76031	268	1.05251	Algibacter
OTU291	0.00135	0.76100	269	1.04956	Syntrophococcus
OTU108	0.00123	0.77123	270	1.05973	Lachnospiraceae Incertae Sedis
OTU424	0.00123	0.77154	271	1.05624	Streptococcus
OTU176	0.00120	0.77451	272	1.05641	Erwinia
OTU119	0.00117	0.77710	273	1.05605	Lachnobacterium

OTU338	0.00116	0.77791	274	1.0533	Micrococcineae
OTU206	0.00106	0.78756	275	1.06249	Paludibacter
OTU182	0.00105	0.78893	276	1.06048	Lachnospiraceae Incertae Sedis
OTU118	0.00104	0.78945	277	1.05735	Burkholderia
OTU57	0.00104	0.78976	278	1.05395	Lachnospiraceae Incertae Sedis
OTU17	0.00098	0.79508	279	1.05725	Escherichia
OTU60	0.00096	0.79778	280	1.05705	Subdoligranulum
OTU89	0.00094	0.79996	281	1.05618	Bacteroides
OTU111	0.00092	0.80186	282	1.05493	Peptostreptococcaceae Incertae Sedis
OTU144	0.00088	0.80648	283	1.05726	Dorea
OTU181	0.00087	0.80664	284	1.05375	Bacteroides
OTU411	0.00081	0.81405	285	1.0597	Faecalibacterium
OTU127	0.00080	0.81495	286	1.05715	Lachnospiraceae Incertae Sedis
OTU91	0.00069	0.82817	287	1.07056	Lactobacillus
OTU285	0.00068	0.82973	288	1.06886	Butyrivibrio
OTU195	0.00067	0.83061	289	1.06628	Pseudoalteromonas
OTU379	0.00067	0.83079	290	1.06284	Roseburia
OTU266	0.00065	0.83282	291	1.06177	Bacteroides
OTU145	0.00063	0.83611	292	1.06231	Afipia
OTU56	0.00062	0.83641	293	1.05907	Delftia
OTU76	0.00062	0.83735	294	1.05666	Lachnobacterium
OTU292	0.00057	0.84278	295	1.05991	Alistipes
OTU168	0.00056	0.84464	296	1.05865	Roseburia
OTU179	0.00056	0.84494	297	1.05546	Ruminococcaceae Incertae Sedis
OTU538	0.00046	0.85925	298	1.06974	Lachnospiraceae Incertae Sedis
OTU319	0.00043	0.86444	299	1.07259	Agrobacterium
OTU360	0.00042	0.86578	300	1.07068	Faecalibacterium
OTU120	0.00041	0.86755	301	1.06931	Micrococcineae
OTU188	0.00040	0.86888	302	1.0674	Lachnospiraceae Incertae Sedis
OTU50	0.00040	0.86920	303	1.06427	Sutterella
OTU387	0.00040	0.86939	304	1.061	Coprococcus
OTU493	0.00038	0.87259	305	1.06141	Lachnospiraceae Incertae Sedis
OTU167	0.00036	0.87483	306	1.06066	Allobaculum
OTU375	0.00036	0.87558	307	1.05811	Pseudomonas
OTU412	0.00035	0.87630	308	1.05554	Sphingomonas
OTU250	0.00033	0.87983	309	1.05636	Paludibacter
OTU409	0.00032	0.88166	310	1.05514	Alkalilimnicola
OTU136	0.00032	0.88268	311	1.05298	Micrococcineae
OTU51	0.00031	0.88342	312	1.05047	Klebsiella
OTU373	0.00029	0.88727	313	1.05168	Sporobacter

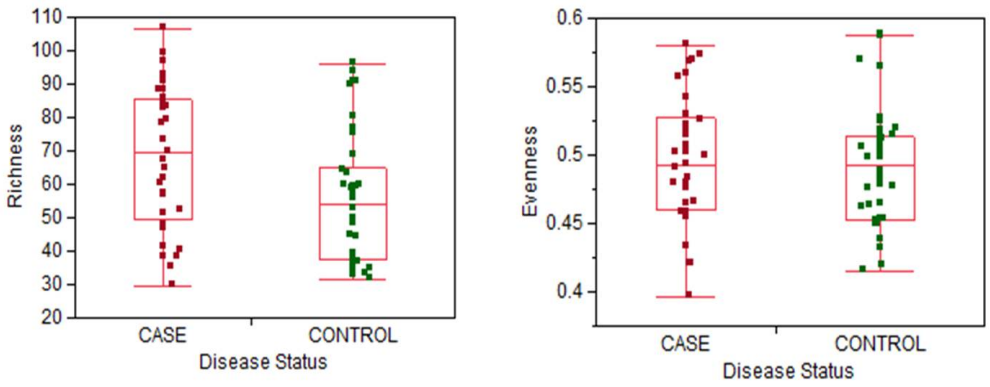
OTU164	0.00029	0.88754	314	1.04866	Faecalibacterium
OTU115	0.00028	0.89031	315	1.04859	Roseburia
OTU260	0.00028	0.89035	316	1.04532	Erysipelotrichaceae Incertae Sedis
OTU491	0.00028	0.89058	317	1.04229	Clostridiaceae 1
OTU97	0.00027	0.89157	318	1.04016	Pseudomonas
OTU408	0.00025	0.89598	319	1.04204	Bryantella
OTU207	0.00023	0.90106	320	1.04466	Succinispira
OTU107	0.00023	0.90113	321	1.04149	Ruminococcus
OTU452	0.00020	0.90578	322	1.04362	Butyrivibrio
OTU341	0.00020	0.90713	323	1.04193	Prevotella
OTU287	0.00020	0.90727	324	1.03888	Anaerovorax
OTU156	0.00019	0.90839	325	1.03696	Lachnospiraceae Incertae Sedis
OTU216	0.00016	0.91636	326	1.04285	Sphingomonas
OTU86	0.00016	0.91719	327	1.0406	Fusobacterium
OTU92	0.00016	0.91754	328	1.03783	Rubrobacterineae
OTU205	0.00013	0.92564	329	1.04381	Erysipelotrichaceae Incertae Sedis
OTU180	0.00013	0.92568	330	1.04068	Roseburia
OTU230	0.00012	0.92648	331	1.03844	Butyrivibrio
OTU196	0.00012	0.92666	332	1.03552	Bacteroides
OTU166	0.00012	0.92794	333	1.03383	Lachnospiraceae Incertae Sedis
OTU139	0.00011	0.93013	334	1.03317	Azonexus
OTU83	0.00011	0.93076	335	1.03078	Dorea
OTU82	0.00010	0.93505	336	1.03245	Roseburia
OTU254	0.00009	0.93617	337	1.03062	Lachnospiraceae Incertae Sedis
OTU304	0.00009	0.93661	338	1.02806	Faecalibacterium
OTU222	0.00009	0.93812	339	1.02667	Prevotella
OTU5	0.00008	0.93979	340	1.02547	Sphingomonas
OTU85	0.00008	0.94221	341	1.0251	Bacteroides
OTU313	0.00006	0.94976	342	1.0303	Enterobacter
OTU233	0.00006	0.94995	343	1.0275	Syntrophococcus
OTU569	0.00005	0.95462	344	1.02955	Erwinia
OTU463	0.00004	0.95591	345	1.02795	Lachnospiraceae Incertae Sedis
OTU345	0.00004	0.95812	346	1.02735	Butyrivibrio
OTU190	0.00004	0.96018	347	1.02659	Ruminococcaceae Incertae Sedis
OTU68	0.00004	0.96091	348	1.02442	Dorea
OTU519	0.00003	0.96198	349	1.02262	Catonella
OTU44	0.00003	0.96309	350	1.02087	Lachnospiraceae Incertae Sedis
OTU71	0.00003	0.96365	351	1.01856	Lachnospiraceae Incertae Sedis
OTU64	0.00003	0.96397	352	1.016	Erwinia
OTU464	0.00002	0.97445	353	1.02414	Marinilabilia

OTU495	0.00001	0.97451	354	1.02131	Streptococcus
OTU248	0.00001	0.97479	355	1.01873	Lachnospiraceae Incertae Sedis
OTU70	0.00001	0.97564	356	1.01675	Sphingobium
OTU160	0.00001	0.97732	357	1.01564	Lachnospiraceae Incertae Sedis
OTU244	0.00001	0.97775	358	1.01325	Prevotella
OTU272	0.00001	0.97876	359	1.01147	Sporobacter
OTU267	0.00001	0.97889	360	1.0088	Parabacteroides
OTU170	0.00001	0.98074	361	1.00791	Bacteroides
OTU303	0.00001	0.98274	362	1.00717	Faecalibacterium
OTU458	0.00000	0.98693	363	1.00868	Roseburia
OTU270	0.00000	0.98704	364	1.00602	Succinispira
OTU393	0.00000	0.98709	365	1.00331	Micrococcineae
OTU400	0.00000	0.98754	366	1.00103	Bryantella
OTU547	0.00000	0.98883	367	0.99961	Subdoligranulum
OTU52	0.00000	0.99158	368	0.99966	Lachnospiraceae Incertae Sedis
OTU69	0.00000	0.99172	369	0.9971	Lachnospiraceae Incertae Sedis
OTU47	0.00000	0.99456	370	0.99725	Succinispira
OTU437	0.00000	0.99660	371	0.9966	Marinilabilia

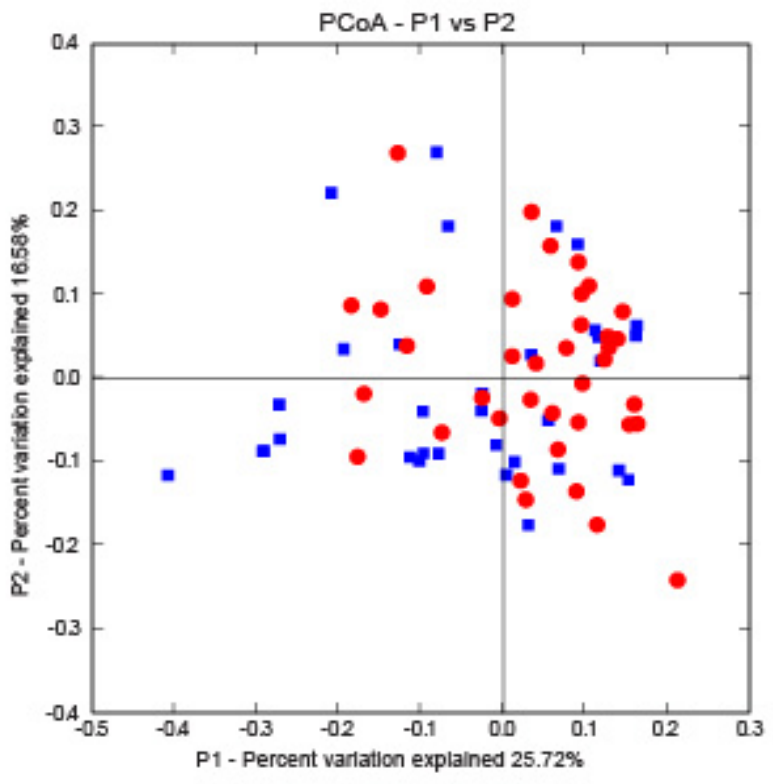
Supplementary Figures



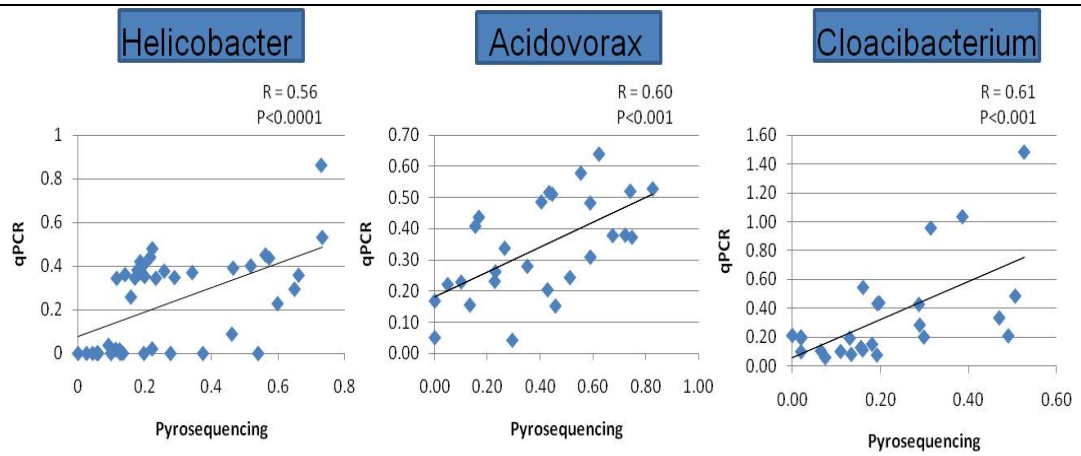
Supplementary Figure 1: Richness (left panel) and evenness (right panel) at the phylum level in cases (n=33) vs. controls (n=38). By the Wilcoxon test, cases had a significantly higher richness (p= 0.0041) than controls, but there was no significant difference in evenness (p = 0.75).



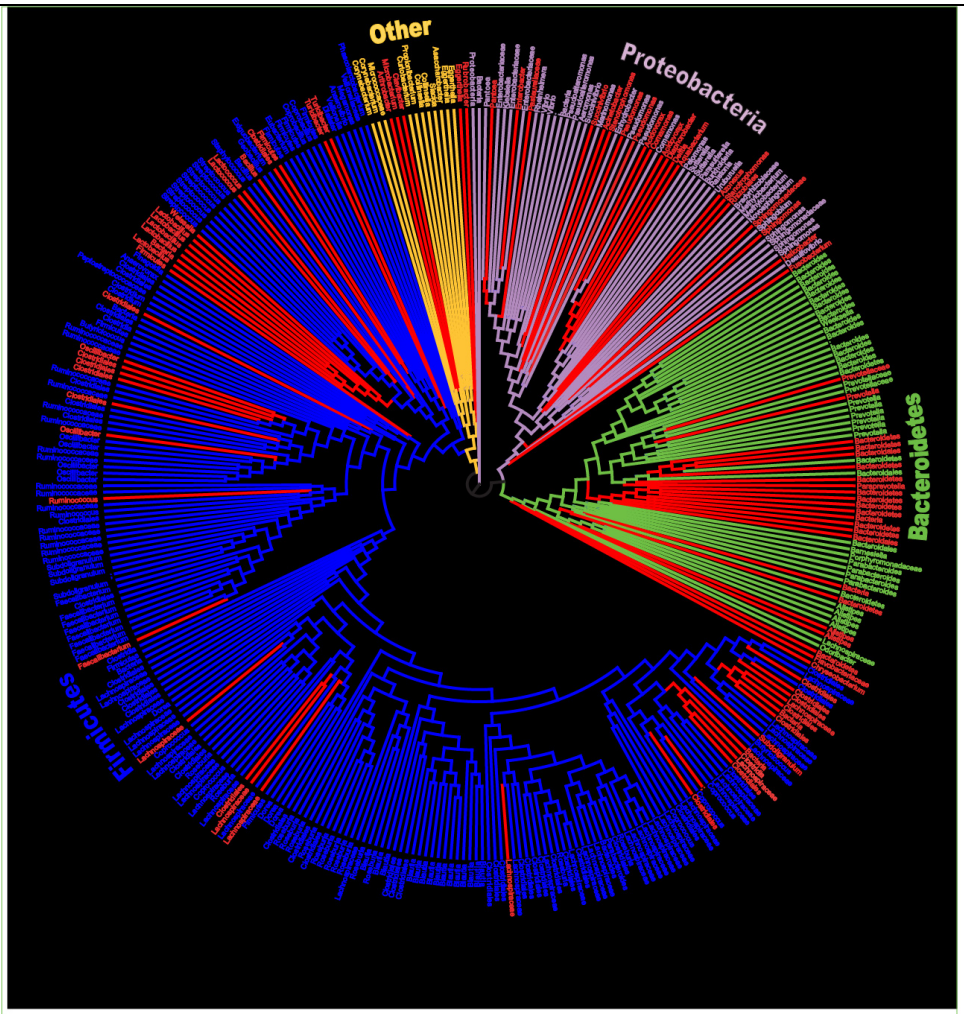
Supplementary Figure 2: Richness (left panel) and evenness (right panel) at the genus level, in cases (n=33) vs. controls (n=38). By the Wilcoxon test, cases had a significantly higher richness (p= 0.0013) than controls, but there was no significant difference in evenness (p = 0.56).



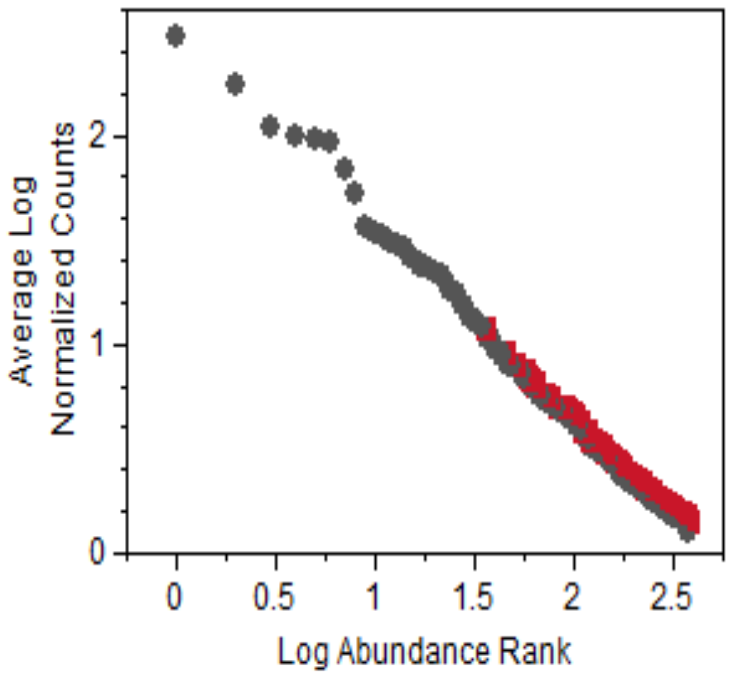
Supplementary Figure 3: PCoA generated from Fast UniFrac analysis on the tree displayed in Figure. 2. (Cases- blue squares; controls- red circles).



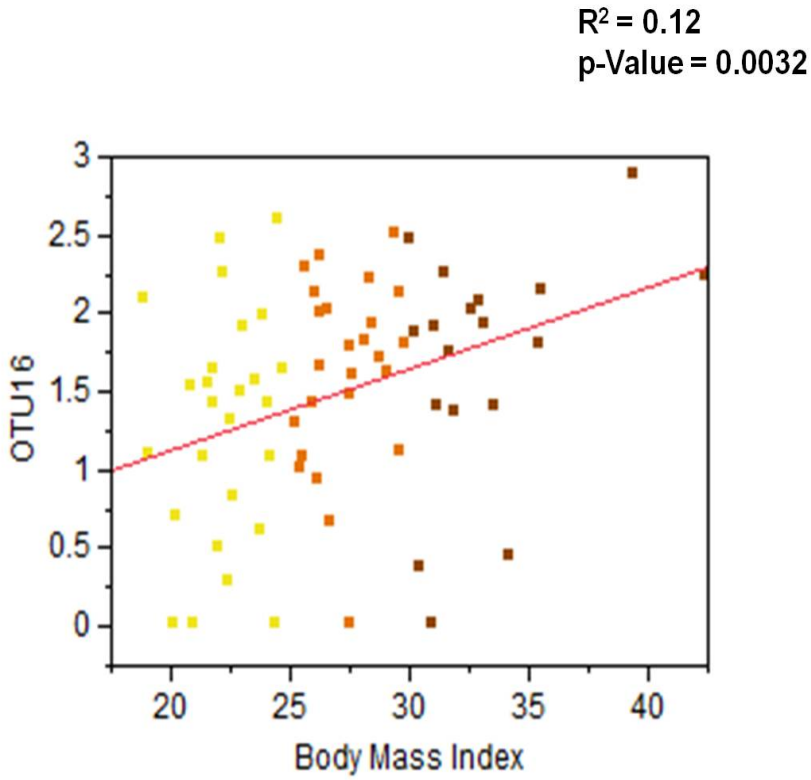
Supplementary Figure 4: Regressions between q-PCR results and results from pyrosequencing data for genera Helicobacter, Acidovorax and Cloacibacterium. We obtained reasonable correlations between the two methods; by linear regression: Acidovorax R= 0.6, p< 0.001; Cloacibacterium R= 0.61, p<0.001 and Helicobacter R= 0.56, p < 0.0001.



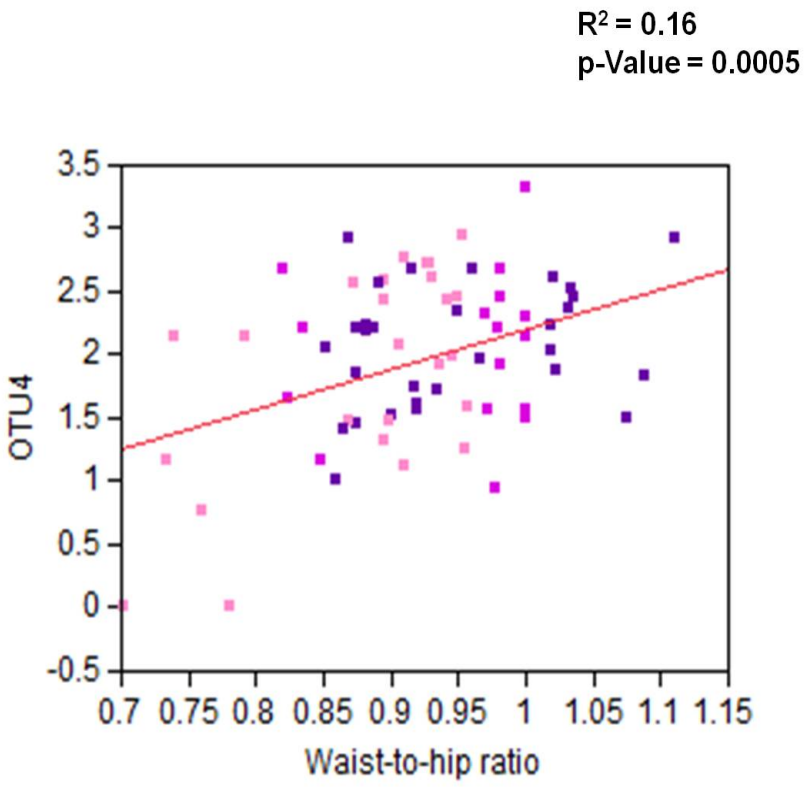
Supplementary Figure 5: Maximum likelihood tree generated from the top 371 OTUs using RaxXML EPA server. Leaf nodes are labeled with the RDP call of the consensus sequence at 80%. Branches are colored red if the OTU was significantly different between case and control and blue if not significant (at 10% FDR).



Supplementary Figure 6: Rank-abundance curve in which the x-axis is the log abundance rank of the top 371 OTUs and the y-axis is the average log normalized sequence count across all samples. The OTU is marked by red squares if the difference between cases and controls is significant at 10% FDR and by black circles if the difference is not significant at 10% FDR.



Supplementary Figure 7: Regressions on log-normalized abundance of OTU16 (top ranking OTU based on regression p-Value) vs. BMI of all samples. Note that after correction for multiple hypothesis testing, this regression is not significant at a 10% FDR threshold (see Appendix A: Supplementary table 6).



Supplementary Figure 8: Regressions on log-normalized abundance of OTU4 (top ranking OTU based on regression p-Value) vs. WHR of all samples. Note that after correction for multiple hypothesis testing, this regression is not significant at a 10% FDR threshold (see Appendix A: Supplementary table 7).

Supplementary References

1. Benjamini, Y. & Hochberg, Y. A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.* **Series B (Methodological) Vol. 57**, 12 (1995).
2. Ley, R.E., *et al.* Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* **102**, 11070-11075 (2005).

APPENDIX B: SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Supplementary Tables

Supplementary Table 1: Wastewater dataset: Spearman's correlations and p-Values between the PCR 16S gene sequence based and whole genome sequence based methods at all taxonomic levels.

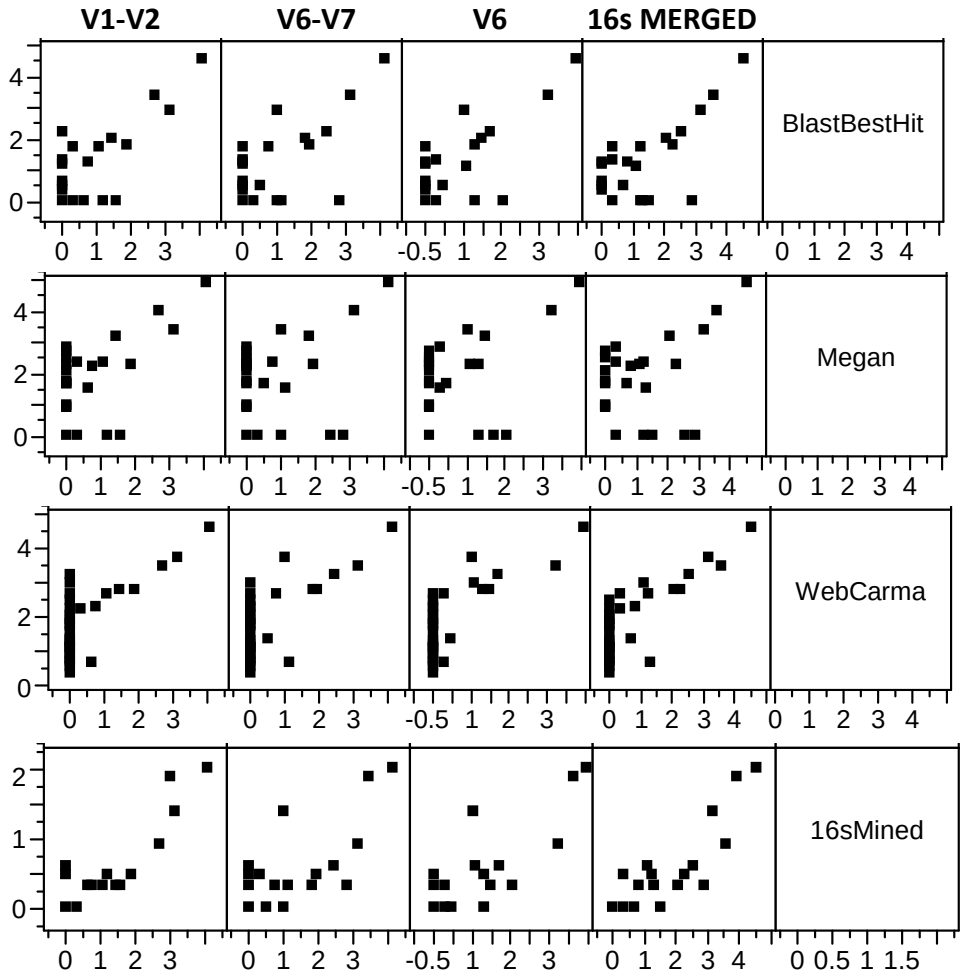
PHYLUM	V1-V2	V1-V2 p-Value	V6-V7	V6-V7 p-Value	V6	V6 p-Value	16sMerged	16s Merged p-Value
BlastBestHit	0.4129	0.0562	0.3655	0.0944	0.3959	0.0682	0.4318	0.0448
Megan	0.3543	0.0894	0.099	0.6455	0.1965	0.3574	0.185	0.3869
WebCARMA	0.5558	0.0002	0.5325	0.0004	0.6091	0.0001	0.6901	0.0001
16sMined	0.639	0.0018	0.6208	0.0027	0.5848	0.0054	0.7766	0.0001
CLASS	**V1-V2**	**V1-V2 p-Value**	**V6-V7**	**V6-V7p-Value**	**V6**	**V6 p-Value**	**16sMerged**	**16s Merged p-Value**
BlastBestHit	0.1613	0.2957	0.0851	0.5827	0.2134	0.1642	0.1126	0.4668
Megan	0.4018	0.0056	0.1584	0.293	0.2625	0.078	0.2508	0.0927
WebCARMA	0.0414	0.7661	0.0372	0.7892	0.1473	0.288	0.0116	0.9335
16sMined	0.7272	0.0001	0.5141	0.0061	0.6347	0.0004	0.7637	0.0001
ORDER	**V1-V2**	**V1-V2 p-Value**	**V6-V7**	**V6-V7p-Value**	**V6**	**V6 p-Value**	**16sMerged**	**16s Merged p-Value**
BlastBestHit	0.2813	0.0152	0.3644	0.0014	0.3213	0.0052	0.3504	0.0022
Megan	0.3964	0.0003	0.3065	0.0057	0.2701	0.0154	0.3698	0.0007
WebCARMA	0.3454	0.0113	0.253	0.0676	0.2455	0.0764	0.3751	0.0057
16sMined	0.7253	0.0001	0.5226	0.0003	0.5712	0.0001	0.7688	0.0001
FAMILY	**V1-V2**	**V1-V2 p-Value**	**V6-V7**	**V6-V7 p-Value**	**V6**	**V6 p-Value**	**16sMerged**	**16s Merged p-Value**
BlastBestHit	0.2079	0.0141	0.2127	0.0119	0.1669	0.0496	0.1625	0.0559
Megan	0.0663	0.4096	0.004	0.9604	-0.0005	0.9948	-0.0121	0.8807
WebCARMA	0.0626	0.5693	0.081	0.4614	0.1114	0.3102	0.0639	0.5611
16sMined	0.5574	0.0001	0.3971	0.0005	0.4308	0.0001	0.6097	0.0001
GENUS	**V1-V2**	**V1-V2 p-Value**	**V6-V7**	**V6-V7 p-Value**	**V6**	**V6 p-Value**	**16sMerged**	**16s Merged p-Value**
BlastBestHit	-0.2742	0.0001	-0.3021	0.0001	-0.3229	0.0001	-0.4554	0.0001
Megan	-0.3514	0.0001	-0.3252	0.0001	-0.3712	0.0001	-0.5329	0.0001
WebCARMA	0.0363	0.6497	-0.0455	0.569	-0.0971	0.2234	-0.1075	0.1774

16sMined	0.3433	0.0001	0.184	0.0262	0.1125	0.1764	0.3553	0.0001

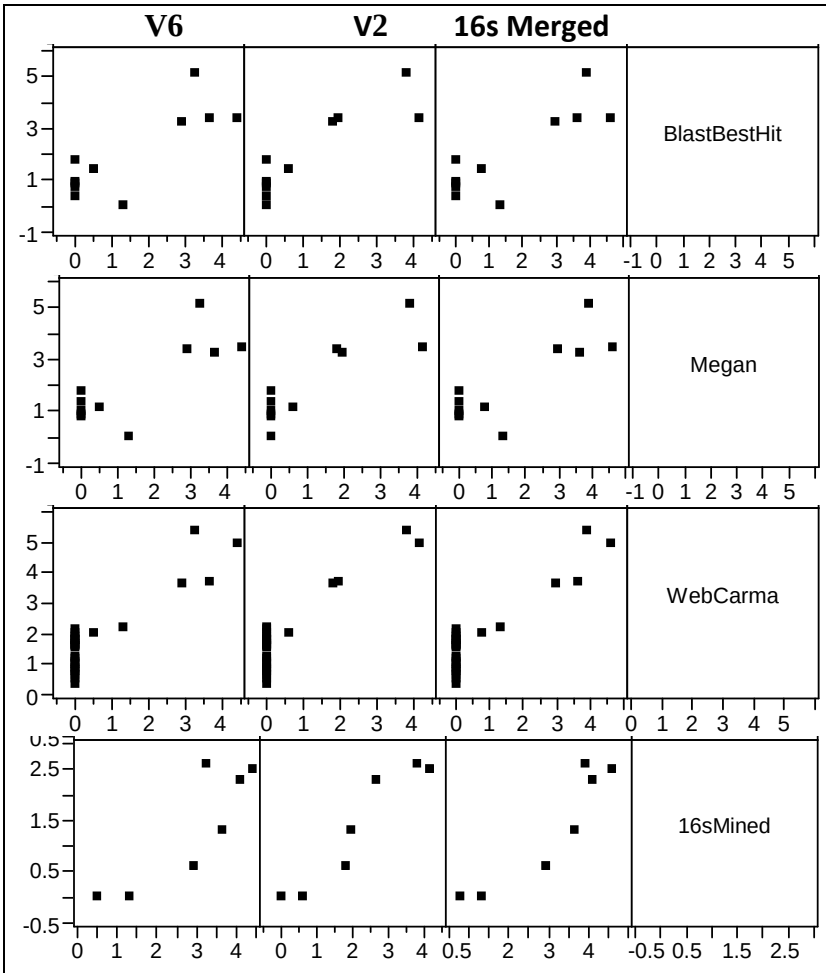
Supplementary Table 2: Human gut microbiome dataset: Spearman's correlations and p-Values between the PCR 16S gene sequence based and whole genome sequence based methods at all taxonomic levels.

PHYLUM	**V6**	**V6 p-Value**	**V2**	**V2 p-Value**	**16sMerged**	**16s Mergedp-Value**
BlastBestHit	0.636	0.0108	0.8166	0.0002	0.6441	0.0096
Megan	0.6289	0.0213	0.8104	0.0008	0.6469	0.0169
WebCARMA	0.6293	0.0001	0.5745	0.0002	0.6301	0.0001
16s Mined	0.7748	0.0408	0.955	0.0008	0.8829	0.0085
CLASS	**V6**	**V6 p-Value**	**V2**	**V2 p-Value**	**16sMerged**	**16s Mergedp-Value**
BlastBestHit	-0.0536	0.7864	-0.0661	0.7381	-0.0592	0.7647
Megan	-0.0488	0.825	-0.0584	0.7912	-0.0558	0.8005
WebCARMA	0.236	0.1276	0.1274	0.4154	0.233	0.1328
16s Mined	0.9174	0.0001	0.943	0.0001	0.9344	0.0001
ORDER	**V6**	**V6 p-Value**	**V2**	**V2 p-Value**	**16sMerged**	**16s Mergedp-Value**
BlastBestHit	-0.1095	0.3967	-0.0298	0.8181	-0.0705	0.586
Megan	-0.4385	<.0001	-0.5327	<.0001	-0.6112	<.0001
WebCARMA	0.1862	0.2208	0.2966	0.0479	0.2354	0.1196
16s Mined	0.892	0.0001	0.8619	0.0001	0.9077	0.0001
FAMILY	**V6**	**V6 p-Value**	**V2**	**V2 p-Value**	**16sMerged**	**16s Mergedp-Value**
BlastBestHit	-0.2677	0.0063	-0.2025	0.0402	-0.2916	0.0028
Megan	-0.2674	0.0133	-0.1743	0.1107	-0.2863	0.0079
WebCARMA	0.2161	0.1239	0.3424	0.013	0.2112	0.1328
16s Mined	0.8734	0.0001	0.8451	0.0001	0.9177	0.0001
GENUS	**V6**	**V6 p-Value**	**V2**	**V2 p-Value**	**16sMerged**	**16s Mergedp-Value**
BlastBestHit	-0.395	0.0001	-0.4939	0.0001	-0.564	0.0001
Megan	-0.4385	0.0001	-0.5327	0.0001	-0.6112	0.0001
WebCARMA	-0.0627	0.6034	-0.1438	0.2315	-0.2103	0.0783
16s Mined	0.5096	0.0003	0.6992	0.0001	0.779	0.0001

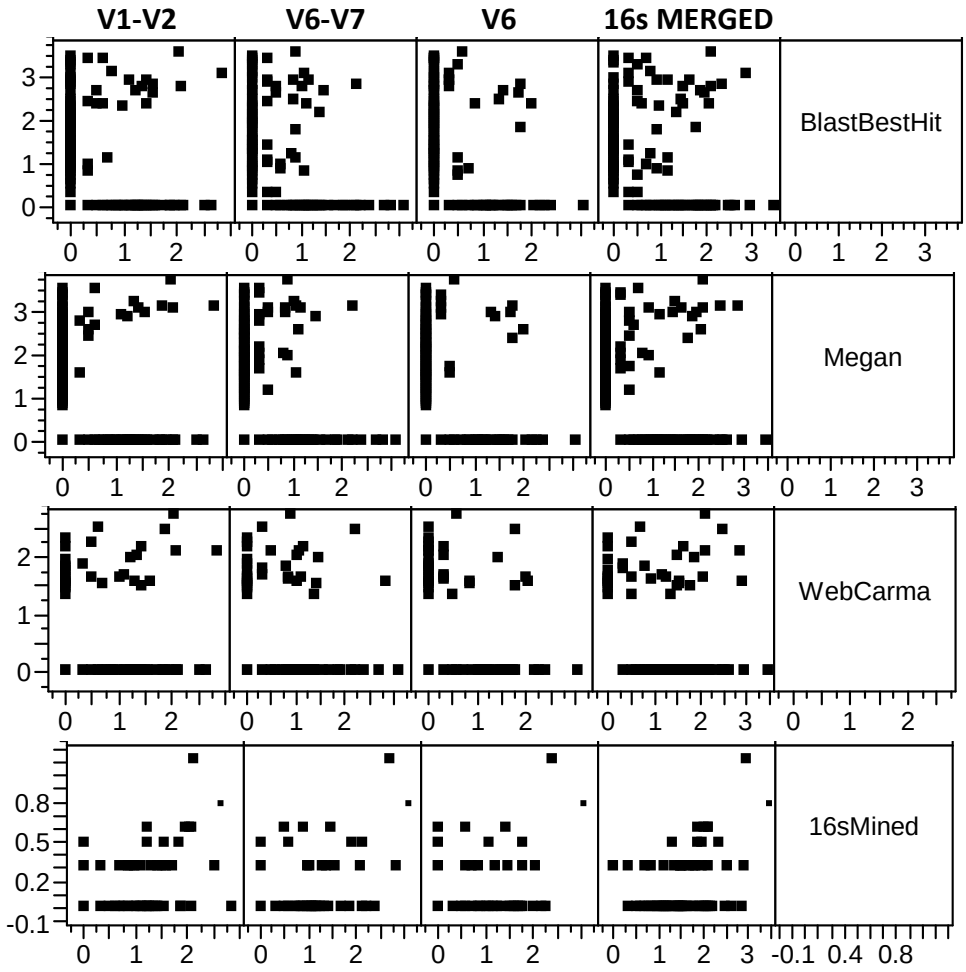
Supplementary Figures



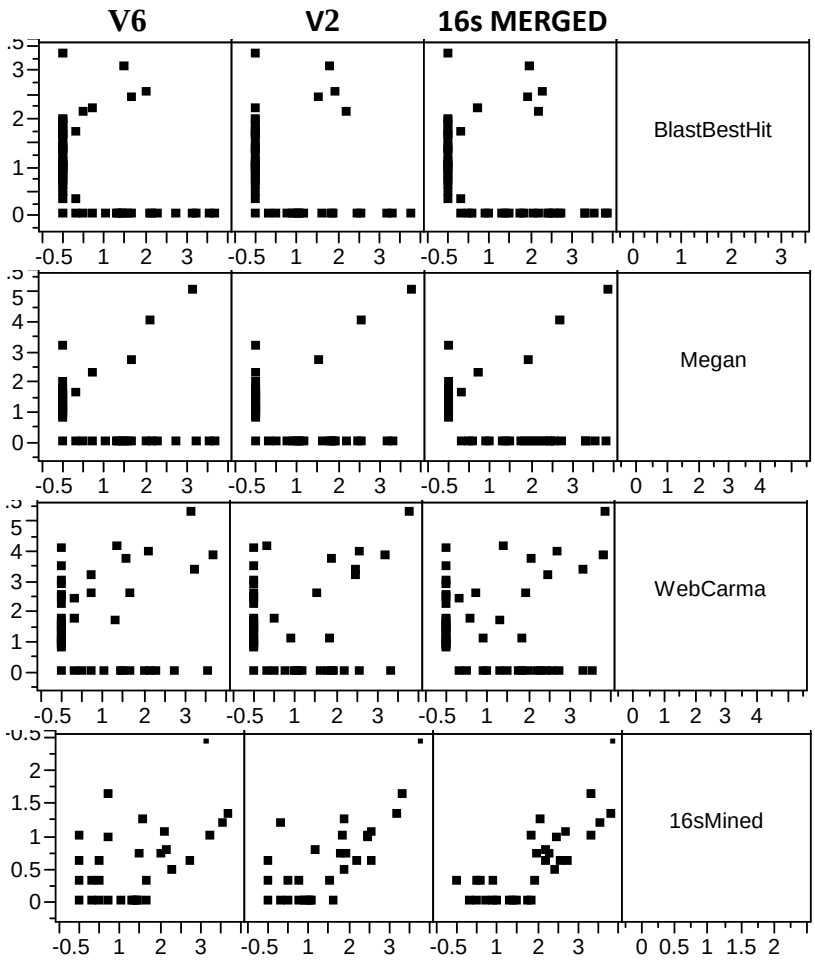
Supplementary Figure 1a: Wastewater dataset: Correlations between the PCR 16S gene sequence based and whole genome sequence based methods at the Phylum level.



Supplementary Figure 1b: Human gut microbiome dataset: Correlations between the PCR 16S gene sequence based and whole genome sequence based methods at the Phylum level.



Supplementary Figure 2a: Wastewater dataset: Correlations between the PCR 16S gene sequence based and whole genome sequence based methods at the Genus level.



Supplementary Figure 2b: Human gut microbiome dataset: Correlations between the PCR 16S gene sequence based and whole genome sequence based methods at the Genus level.