Characterization of Microbial Populations in Aquatic Environments

Methodology and Application of Metagenomics for the Characterization of Bacterial Populations in Aquatic Environments

By Yasser Salama, HBSc

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University
Master of Science (2016)
Hamilton, Ontario (Department of Biology)

TITLE: Methodology and Application of Metagenomics for the Characterization of Bacterial Populations in Aquatic Environments
AUTHOR: Yasser SALAMA (McMaster University)
SUPERVISOR: Dr. Brian GOLDING
NUMBER OF PAGES: ix, 88

Abstract

Metagenomics is a culture-independent framework for deciphering the complexity of biological communities, often with a focus on microbial communities in a specific environment. The applicability of this approach is widespread due to the ubiquity and presence of unculturable microbes in many environments which can only be investigated using culture-independent methods. With advances in DNA sequencing and the introduction of high-throughput sequencing technologies, studying microbial life as communities has become more accessible. However, the breadth of data generated dictates that computational processing steps must be in place to analyze the data. Due to the large diversity in computational and bioinformatic steps possible for metagenomic data, differences in methods of analysis can lead to discordant interpretations of results. The performance of different metagenomics methods must therefore be assessed to establish the effect on the interpretation of results. Taxonomic classification is an integral step in metagenomic analysis and many tools exist for this purpose. To determine which tools are better suited for particular types of metagenomic data, a comparative analysis of performance was conducted for numerous tools. The findings suggest that hybrid programs may have the best performance and warrant further investigation. Programs such as CLARK, KRAKEN, and MEGAN also performed well and are suitable for metagenomic analysis. Utilizing these methods, investigation into the bacterial populations of four freshwater beaches was examined. Bacterial communities in beach waters and sands were more distinct in terms of taxonomic composition than communities in different lakes. Functional capacity was stable between beach habitats, although enrichment of anaerobic and stress related genes in the sand suggests that this is a relatively harsh environment. The detection of sequences belonging to pathogens in the sands of these beaches also has implications for public health and warrants changes in water quality monitoring procedures.

Acknowledgements

Firstly, I want to thank all the members of the Golding and Evans labs, past and present, whom I have had the pleasure of meeting. I've learned a tremendous amount from everyone, especially Wilson who I bugged almost daily when I first started. I also have to thank Shahrokh, Caitlin, Daniella, and Nick who have made the long days in the lab feel shorter and have indulged and joined me in complaining about LaTeX and its nonsensical error codes (but it makes up for that in every other way). You all have been a big part of my enjoyment the lab and have been a great help whenever I had questions.

I also want to acknowledge Dr. Schellhorn and the members of the Schellhorn lab who I worked closely with, Mahi and Sakis, for providing excellent guidance, data, and lots of interesting questions to address in my research; thank you. Importantly, I have to thank my supervisor Brian - you have been a wonderful supervisor and really gave me a strong passion for bioinformatics and computational biology. You made learning new computational skills fun, challenging, and memorable. Your patience in working with me, who had very little computational experience initially, has been much appreciated. You have taught me a lot, and I appreciate your support in pursuing my interests both in and out of the lab.

Lastly, I want to thank all my family and friends who have been extremely understanding and supportive. Thank you to my parents especially who have been supportive of my decisions and continue to motivate me to pursue my interests. Thank you to Kerry for being so wonderful and for always being there for me. I would not have been able to manage without the support of you all.

Contents

List of Tables

List of Figures

Declaration of Authorship

I, Yasser SALAMA, declare that this thesis titled, "Methodology and Application of Metagenomics for the Characterization of Bacterial Populations in Aquatic Environments" and the work presented in it are my own. My contributions to the work in the main text are as follows:

- Chapter 1
  - Research and writing
- Chapter 2
  - Conception of research idea and methodology
  - Execution of methods and development of results
  - Analysis and discussion
- Chapter 3
  - Development of research questions
  - Methodology and computational processing
  - Analysis and discussion

Chapter 1

Introduction

1.1 Microbial Life

Microbial life is ubiquitous in every part of the biosphere. Prokaryotes alone are estimated to number $4\text{–}6 \times 10^{30}$ cells globally [1], and their inhabitance ranges from nutrient-dense and optimal environments such as soils and aquatic habitats to extreme geochemical and nutrient-deficient conditions such as hydrothermal vents [2], anoxic marine sediments [3], and acidic mine drainages [4]. It is quite evident that the ubiquity of these microscopic organisms in such contrasting environments is underpinned by a large diversity of biochemical and metabolic processes that can mediate survival and proliferation. The vast and complex biological repertoire microorganisms possess facilitates their resilience, adaptability, and diversity in numerous environments. In addition to the large genetic diversity harbored by microorganisms, their impact on nutrient cycles and as symbionts in a multitude of ecosystems exemplifies their global importance. Microorganisms harbor the largest pools of nitrogen and phosphorus of any organisms, and harbor an estimated cellular carbon of $350\text{–}550\,\mathrm{pg}$, of which the upper range is as much as the carbon stored in plants [1]. Microorganismal symbiotic relationships with plants, such as the nitrogen-fixing rhizobial association with legumes, also contribute largely to plant growth and nutrient uptake, and impact natural ecosystems and agricultural processes [5]. From the beneficial microbiota of humans [6–8] which can impact energy consumption and nutrient uptake [6], obesity and fat storage [7], and other phenotypes [8], to the pathogenic and infective agents which cause disease, microorganisms and especially bacteria have a dramatic impact on human life as well. Microorgainsms inhabit all regions of the human body and outnumber human cells by an order of magnitude [8]. Bacteria have also been a significant source of antibiotics and pharmacologically relevant compounds which are in use today, further highlighting their importance [9]. The ubiquity and immense human and gloabl impact of microbes is a major driver for understanding these organisms and warrants the study and investigation of these organisms.

1.2 Microbial Communities: Application of Metagenomics

Despite the best efforts of microbiologists to study and understand microorganisms, it has become clear that studying single organisms in isolation for the purpose of understanding the complex microbial communities that exist in nearly every environment does not allow for the full characterization of microbial interactions [10]. Traditional culture-dependent methods which attempt to recreate optimal growth conditions of microbes and subsequent assessment of morphological, physiological, and genetic properties of microorganisms has provided insight into many aspect of microbial life and process regulation. However, these methods have been deemed inadequate for describing the complex interactions which exist between numerous microbes in natural environments, which are often the main concerns of microbial ecologists [11]. Additionally, it has become evident that investigating microbial populations as they exist in the environment would be dependent on recognizing the large proportion of unculturable microorganisms in these habitats. This was initially recognized due to the large discrepancy between microscopically distinct morphologies observed and the number of distinct colonies that could be grown on a nutrient rich medium [12]. This was a sign that the standard conditions presented for culturing microorganisms was a major bottleneck and that there is a distinct and large diversity of organisms that could not be easily cultivated. This finding was verified in numerous habitats including soil [13], sediments [14], and many aquatic habitats [14, 15], where up to 99% of microorganisms can not be captured through culture-dependent methods [16].

The introduction of culture-independent methods has therefore been a major development in studying microbial life. Culture-independent methods bypass the requirement of simulating specific conditions to grow microbes and allow for the direct assessment of microbes in their environment [16]. This approach circumvents the hindrance imposed by traditional culture-dependent methods and has been extremely valuable in uncovering the previously underestimated diversity of microbial life. Culture-independent methods targeting bacterial populations largely relied upon the small subunit (SSU) ribosomal RNA (rRNA) initially [16]. Early methods used the extraction of DNA from the environment in a non-selective fashion, and either direct detection of 5S or 16S sequences using electrophoretic methods [17, 18] or shotgun cloning into an appropriate vector [19], subsequent screening for rRNA sequences using probes of interest, and sequencing. Developments of these preliminary methods involved the addition of PCR or reverse transcribing of target sequences to selectively amplify the region of interest [16, 20]. This greatly simplified the process by permitting the direct enrichment of 16S sequences from microbial DNA assemblages using primers complementary to highly conserved regions of the gene and avoids the requirement of cloning. Early studies which adopted this method had relatively good outcomes and showed early signs of discovery of novel 16S sequences, supporting the notion of many undiscovered organisms in different environments [20, 21].

Methods which rely upon enrichment of 16S or other phylogenetically potent signals from DNA are termed amplicon sequencing. Amplicon sequencing can be applied for bacterial [22] and archaeal identification using 16S sequences [23], and eukaryotic identification by sequences such as the 18S SSU rRNA gene or the internal transcribed spacer (ITS) region [24, 25]. However, the largest group targeted for analysis are bacteria, and this is reflected in the growth of bacterial SSU rRNA sequences in databases (Figure 1.1) [26]. The selective amplification of different sequences is advantageous in that different primers can be used to target these different populations (e.g. 16S primers specific for bacteria or archaea), but also imposes a bias in the detection of the community of interest. Although primer sets for PCR of 16S or other sequences have been developed with the purpose of capturing different microbial populations, preferential amplification of certain sequences can skew the capture of the community structure. A contrasting approach to the amplicon sequencing method is one that utilizes whole genome information from microbial communities. The whole genome shotgun (WGS) metagenomic approach can therefore identify information beyond the taxonomic composition and also bypasses any biases that are introduced through amplification. Originally, metagenomic studies required the extraction of the collective microbial DNA from an environment of interest followed by cloning into an appropriate vector and then traditional Sanger sequencing. The use of metagenomics has yielded insight into both the taxonomic and functional classification of microbial communities in numerous environments. Early works investigated microbial communities in the Sargasso sea [27], acidophilic biofilm [28], whale falls, acid mine drainages, and soils [29]. These works were successful in identifying important and unique microbial features representative of the environments, and even the detection of novel genes (e.g. rhodopsin-like photoreceptors in the Sargasso sea [27]).

The shotgun metagenomic approach has been dramatically improved upon by the introduction of high-throughput sequencing (HTS) and parallelization (e.g. 454 [30] and Illumina [31]), which much like the progression of 16S sequencing approaches with the introduction of PCR, avoids the cloning step and permits direct sequencing of DNA extracted from an environment [32]. Although early work by Venter et al. [27] and Tringe et al. [29] revealed significant gaps in our microbial understanding in unique environments, the limited throughput of the cloning and traditional sequencing narrowed the applicability of this method to low complexity microbial communities or only permitted the investigation of the most abundant organisms in a given environment [33, 34]. HTS in combination with whole genome extraction of microbial communities has therefore provided an avenue for a much more comprehensive exploration of microbial communities and the ability to examine taxonomic and metabolic properties of environmental microbial populations [34]. Although HTS approaches typically produce sequence reads (reads) which are much shorter than traditional methods, Liu et al. [35] showed that these reads are sufficient for investigating microbial communities. Additionally, with the observation that a greater number of short reads would be more beneficial than less longer reads, even in 16S sequences, the integration of HTS in amplicon-based and metagenomic analysis was further solidified [35].

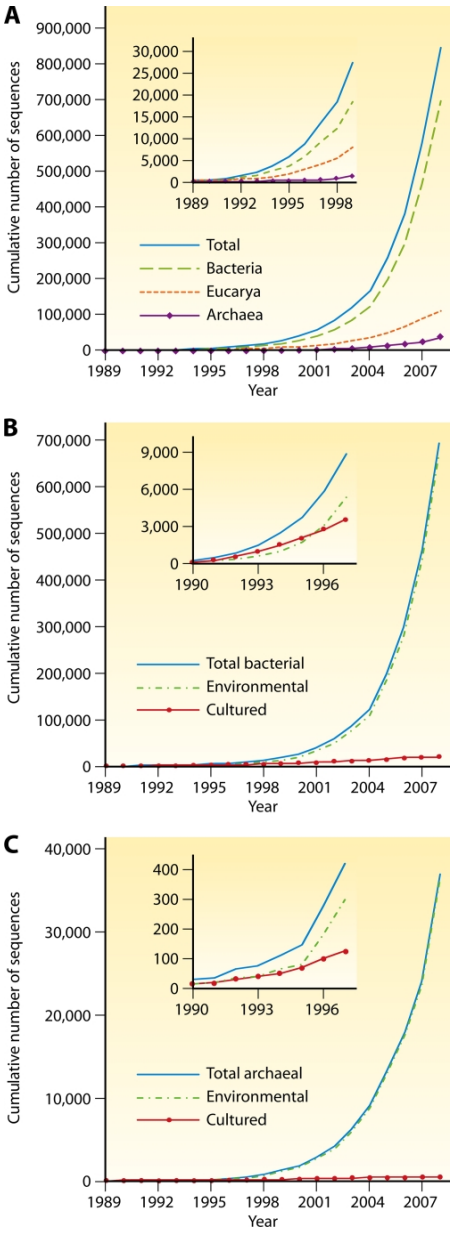


FIGURE 1.1 Accumulation of SSU rRNA sequences over time. This figure is from work by Pace [26] and is data obtained from the SILVA database of SSU rRNA sequences as of 2008. In (A), total, bacterial, eukaryotic, and archaeal sequences are plotted. In (B), bacterial sequences obtained from culture and environmental settings are plotted. In (C), archaeal sequences obtained from culture and environmental settings are plotted. In recent times, environmental sampling has provided a large influx of new SSU rRNA sequences to light, and bacterial sequences are the major driving force of this discovery.

WGS metagenomic methods typically produce larger amounts of data than amplicon-based methods and are more accurate in capturing bacterial populations [36]. Amplicon-based methods typically have lower taxonomic resolution than WGS methods [37] and are prone to amplification bias, potentially skewing the representation of the community towards organisms with sequences more similar to the primers used and underrepresenting taxa with more diverged sequences. Indeed, numerous studies have identified gaps in microbial understanding based on amplicon-based methods, especially for 16S-based analyses of bacterial communities. Work by Ranjan et al. [36] on a human fecal sample identified a greater sensitivity in detecting organisms using a shotgun approach compared to 16S sequencing, and therefore greater accuracy in determining diversity [36]. Furthermore, Eloe-Fadrosh et al. [38] found that applying commonly used primer sets for bacterial identification would not be able to identify at least 9.6% of 16S sequences from metagenomic datasets, and this value increased up to 22% depending on the primers used [38]. Their analysis also suggests that combining multiple primer sets to increase the bacterial capture does not alleviate this problem and numerous SSU rRNA sequences could not be retrieved still. Their work identifies members of a recently described group termed the Candidate Phyla Radiation (CPR) [39] as being the most likely targets to be missed using 16S-based analysis. Brown et al. [39] have determined that members of the CPR are a prominent group of bacteria, making up more than 15% of all bacterial taxa and due to divergent 16S sequences, are not easily detectable using 16S-based approaches [39]. The members of the CPR are environmentally diverse and seem to harbor a unique genetic and biological makeup, highlighting the importance and benefits of shotgun metagenomic approaches for the discovery of rare or unknown organisms.

1.3 Metagenomic Methods and Challenges

Many challenges exist in culture-independent and metagenomic methods. Initial and current applications of amplicon sequencing face the problem of biased amplification due to many factors as mentioned before. The largest factor is likely the use of specific primer sets which may preferentially amplify certain sequences over others [22]. This can be used advantageously to selectively probe for bacterial, archaeal [23], or eukaryotic sequences (e.g. 18S, ITS1) [24, 25], but also introduces the problem of not capturing the microbial population of interest completely, since previously unknown organisms may have sufficiently diverged sequences from the primers. As well, different sequencing platforms may give rise to different interpretations of the same datasets [36, 40], and comparisons of 16S amplicon sequencing and WGS metagenome sequencing of bacterial populations may be incongruent in some details, although general properties of the communities are the same in both methods [41].

Although the aforementioned challenges associated with metagenomic methods are not exhaustive, one important obstacle is the diversity of bioinformatic processing

tools developed for metagenomic analysis [42]. WGS metagenomic approaches produce a large amount of data and are amenable to a variety of processing and analysis steps. Therefore, processing of metagenomic data can vary greatly depending on the goals of the researcher and the type of data available. For most datasets, adapter- and quality-trimming are usually important first steps to remove erroneous sequences from the reads. Read merging for paired-end reads may be a subsequent step for amplicon and metagenomic sequencing depending on the method of library preparation. With sufficient coverage in metagenomic data, assembly is possible and can be used to build genomes from environmental datasets, or even recover complete SSU rRNA sequences. Importantly, however, taxonomic classification is a major step in some amplicon sequencing and metagenomic approaches. This is often a main goal of most metagenomic studies. Further analysis of metagenomic data can include functional annotation of sequences to characterize the functional capacity of the microbial community and downstream statistical analyses to assess patterns of ecologically relevant metrics and changes in microbial community structures. The numerous steps that can be undertaken in metagenomic data analysis is a challenge since it limits the efforts of standardization of analysis. This poses a problem for the analysis and replicability of experiments. To further the understanding of bioinformatic processing on metagenomic data and to push forward the standardization of methods, Chapter 2 will focus on assessing performance of numerous taxonomic classification methods, an integral step in metagenomic analysis.

1.4 Metagenomics in Aquatic Environments

Chapter 3 of this thesis will focus on metagenomic analysis within the context of aquatic environments and, more specifically, of freshwater beaches. Freshwater beaches are known to harbor complex microbial communities consisting of microbial eukaryotes [43], archaea [44], bacteria, and viruses [45, 46]. Metagenomic analysis of microplankton communities in a freshwater lake in the U.S. has revealed relatively stable functional gene content over time, although the communities were phylogenetically distinct from other environments such as soils and marine environments [47]. Analysis of harmful cyanobacterial blooms in freshwater environments reveals again that functional capacity is more stable than taxonomic changes between different blooms [48]. Metagenomic analysis has also revealed the presence of highly abundant aquatic Actinobacteria in freshwater and associated actinorhodopsin complexes [49, 50]. Metagenomics has also been applied to understanding breakdown of the toxin microcystin in Lake Erie [51], although there have been few metagenomic analyses of this lake or Lake Ontario as well.

The focus of Chapter 3 will therefore be on characterizing microbial populations in four freshwater beaches across both Lake Erie and Lake Ontario, and identifying predominant bacterial taxa and their functions. This analysis will also serve to identify trends in bacterial populations between beach waters and sands. Since freshwater

beaches are popular locations for recreational swimming, investigating these areas using a metagenomic approach is useful to understand the exposure of the public to the microbial communities present in these regions.

Chapter 2

Comparative Analysis of Taxonomic Classification Programs

2.1 Introduction

A major step in understanding collective microbial populations and elucidating the microbial diversity of environmental populations through culture-independent metagenomic approaches requires the employment of sensitive, specific, and time-sensible taxonomic classification methods [52]. Amplicon sequence based studies often may not be concerned with determining the taxonomic origin of the reads obtained, and instead focus solely on "binning" reads into clusters of similar sequences. This allows for the investigation of read clusters as operational taxonomic units (OTUs), with each OTU representing a distinct taxonomic unit. In these cases, taxonomic classification is not required. Other amplicon based studies may perform binning, but also want to taxonomically classify these reads to label the OTUs. This type of analysis would require the employment of taxonomic classification. Shotgun metagenomic studies do not have the ability to cluster reads since sequences may originate from anywhere in the genome (barring metagenomic islands due to abnormal sequence composition [53]), and therefore must rely on taxonomic classification of sequences. Taxonomic classification is therefore an integral step of shotgun metagenomic analysis and is required for the labelling of sequence reads to provide information regarding the taxonomic constituents of a sample. This chapter will focus on taxonomic classification within the context of shotgun metagenomic analysis.

In general, taxonomic classification programs possess an underlying similarity-based or composition-based algorithm, although some programs implement elements of both into a hybrid method. Alignment-based methods rely on the alignment of reads to reference sequences and using sequence similarity to assign a taxonomic classification [54]. Composition-based methods exploit the phylogenetic signals present in sequence composition information [55] such as GC content, codon usage, and oligonucleotide (k-mer) patterns to classify metagenomic reads [56–60]. There have been few studies examining the comparative efficiency of these classification methods. Additionally, there is a large range of programs that employ these methods with variations

(e.g. incorporation of interpolated Markov models [58], taxonomy trees [54, 59, 60], and support vector machines [56]). Furthermore, different classification programs may be optimized for specific inputs [61], thus different sampling environments, community complexities, or metagenome qualities also may demand different classification methods for optimal results.

To assess the comparative performance of alignment-based and similarity-based classification methods and to determine the effect of different metagenome inputs on performance, metagenomes were simulated under varying conditions and classification performance was quantified. The programs for analysis were selected on the basis of recency (release or update), citation and usage, and availability (Table 2.1).

TABLE 2.1 Taxonomic Classification Programs

Program	Method	Notes	Reference
blastn + MEGAN	Similarity	Nucleotide alignment, LCA	[54]
blastx + MEGAN	Similarity	Protein alignment, LCA	[54]
Taxator-tk	Similarity	Nucleotide alignment	[62]
Kraken	Composition	Taxonomy tree, k-mer, LCA	[59]
LMAT	Composition	Taxonomy tree, k-mer, genome association	[59]
CLARK	Composition	Unique k-mers	[63]
Phymm	Composition	IMM	[58]
PhymmBL	Hybrid	IMM, nucleotide alignment	[58]

MEtaGenome ANalyzer (MEGAN) is a program commonly used for taxonomic classification [54]. Input is provided to MEGAN in the form of blast(N/X) output, and a lowest common ancestor (LCA) algorithm is used to assign a read to a taxon which encompasses a subset of taxa that a read significantly aligned with. Taxator-tk is an alignment-based method which utilized nucleotide alignment followed by phylogenetic inference of the read origin based on subsegments of local alignment which are weighed based on similarity. Kraken [59] and LMAT [60] are both compositional methods for taxonomic classification, and utilize k-mer content along with taxonomic trees for assignment. LMAT extracts each k-mer in a read and maps it to a taxonomic tree of reference genomes. Kraken finds all taxa which are the LCA of each k-mer and weighs those taxa by the number of k-mers from a single read that map to it. These two programs differ in database structure as LMAT records all genomes associated with a k-mer, while Kraken records only the LCA for a k-mer. CLARK is similarly a k-mer based method, however, k-mers that are common between target sequences (references) are removed from the database and only unique k-mers are retained. Lastly, Phymm is a compositional method which utilizes interpolated Markov models to characterize nucleotide distributions of reference species [58]. Then, by comparing the nucleotide distribution of the query, a score can be assigned to a taxa and a classification can be made. PhymmBL supplements the IMM method of Phymm by incorporating BLASTN results to assign a read [58].

2.2 Methods

2.2.1 Read Simulations

Investigating performance of taxonomic classification programs requires the generation of simulated metagenomes so that the origins of reads are known and may be compared to the classifications assigned by each program. Since the results from this project were intended to be utilized for the analysis of aquatic environments, simulated metagenomes were generated with taxonomic profiles representative of previously analysed aquatic metagenomes (one Fifty Point, one Nickel Beach, and nine Cootes Paradise samples, provided by Mohammad Mohiuddin). All bacterial species identified from these previous samples were sorted by decreasing frequency of detection in the samples, and then by decreasing average relative abundance in the samples. Scripts were then used to generate three taxonomic profiles representing low complexity (one dominant species), medium complexity (multiple dominant species), and high complexity (no dominant species) populations (as defined by Charuvaka and Rangwala [64]). Dominant species were randomly selected from the top 5 of the list of species and given a simulated relative abundance around 10. The rest of the species were randomly selected from the top 250 species with even relative abundances yielding a total of 100 taxa in each taxon profile. MetaSim [18] was used to simulate 200,000 100 bp paired-end Illumina reads using each of the three taxonomic profiles with high, realistic, and no sequencing error, yielding nine metagenomes in total.

Metagenomes were also generated to simulate cases where reads originate from non-reference sequences. The ability to classify reads to the correct taxonomic lineage when a reference sequence is not present in the database is important to assess since this is often the case in metagenomic datasets [16]. Non-reference sequences were generated using MetaSim's "evolve" function. Three simulations were generated to act as technical replicates and only differed in which species were selected for random read extraction.

2.2.2 Classification

Classification was conducted using default settings for all programs. For MEGAN, reads were subjected to blastn (version 2.2.29+) [65] against the nt database or DIAMOND blastx (version 0.7.8.57) [66] against the nr database. All other programs were assessed against the RefSeq bacterial database. For programs with rank-flexible assignments, classification was always chosen at the species level.

2.2.3 Performance Quantification

In order to quantify performance of the taxonomic classification accuracies of each program, measures of classification sensitivity and specificity were adapted from McHardy et al. [56] and Baldi et al. [67] and used as follows:

$$specificity = (\sum_{i=1}^{N} \frac{c_i}{a_i}) \frac{1}{N}$$

$$sensitivity = (\sum_{i=1}^{N} \frac{c_i}{t_i}) \frac{1}{N}$$

where: c_i is the number of correctly assigned reads to clade i
a_i is the total number of reads assigned to clade i
t_i is the number of reads truly belonging to clade i
N is the number of clades

c_i is considered as the true positives for clade i, a_i includes the true positives and false positives wrongly assigned to clade i, and t_i includes the true positives and false negatives belonging to clade i but assigned to another clade. In general, sensitivity acts as a measure of how many of the reads that belong to a specific taxon were captured. This is an indication of how well a program is able to capture the taxonomic composition. A low sensitivity signals that many of the reads truly belonging to a specific clade were not identified. In comparison, specificity measures the accuracy of classifications made by looking at the proportion of classifications at a specific taxon that are correct. A low specificity indicates that many of the reads assigned to a taxon were incorrectly assigned. The measures of sensitivity and specificity were used for each program and each metagenome at taxonomic levels ranging from phylum to species.

2.3 Results

Except for the composition-based Phymm program, all programs exhibited consistent and almost perfect specificity across all taxonomic ranks when no sequencing error was introduced (Figure 2.1). Specificity did not differ between community complexities, and only slight variations in specificity were seen for PhymmBL at less specific taxonomic ranks when sequencing error was introduced. This indicates that these programs

are usually correct when they make a classification to a certain clade. However, taxonomic classification programs varied in sensitivity in an error-dependent manner (Figure 2.2). All programs showed higher sensitivity at less specific taxonomic ranks, with typically worsening sensitivity as ranks became more specific towards the species level. Most programs had their lowest sensitivity at the species level, although PhymmBL, Kraken, and CLARK maintained similar sensitivity at the species level as they did across all other ranks. With no sequencing error, PhymmBL, Kraken, and CLARK performed with perfect sensitivity at all taxonomic ranks. This was followed closely by MEGAN using blastn, although slight decreases in sensitivity were seen when using the RefSeq database in place of the nt database at the species level. LMAT performed with relatively good sensitivity, but was worse than PhymmBL, Kraken, CLARK, and Megan using blastn at all taxonomic ranks. MEGAN with blastx, Phymm, and Taxator-tk exhibited poor sensitivity in cases with no sequencing error. Sensitivity did not seem to vary clearly between community complexities as most programs retained equal sensitivity across low, medium, and high complexity metagenomes. As sequencing error increased however, sensitivity decreased for the majority of programs. However, even with the introduction of sequencing errors, all composition-based programs had greater than 80% sensitivity, which was matched only by MEGAN using blastn and PhymmBL. The decrease in sensitivity can be visualized in Figure 2.3. Although MEGAN using blastx had poor sensitivity with no sequencing error, an increase in error did not translate to as large a decrease in sensitivity as the other programs. This was also true for LMAT. The greatest declines in sensitivity were seen in Phymm, Taxator-tk, CLARK, and Kraken, while MEGAN using blastn and Phymmbl had intermediate drops in sensitivity.

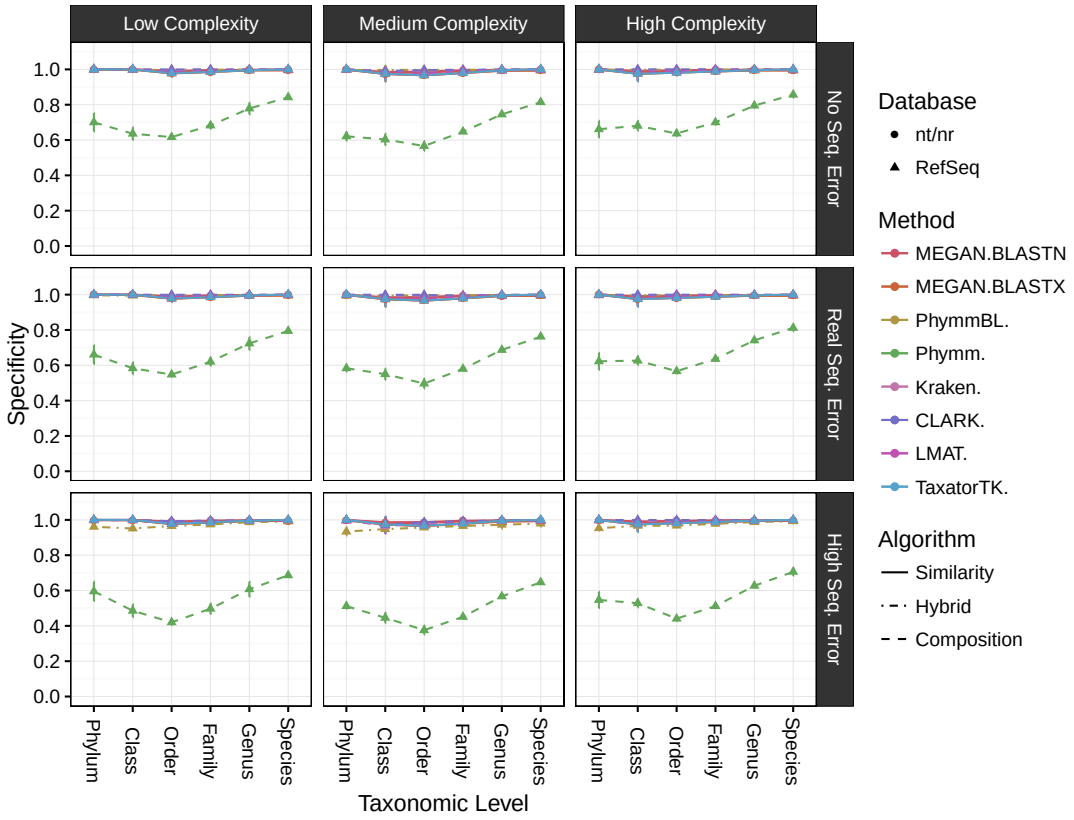


FIGURE 2.1 Specificiy of taxonomic classification programs. Specificity of taxonomic classification programs was measured in triplicate at varying levels of community complexity and sequencing error. Specificity was examined at six taxonomic levels. The database used as a reference, the method, and the underlying algorithm governing each method are indicated. Error bars indicate the standard deviation of the mean.

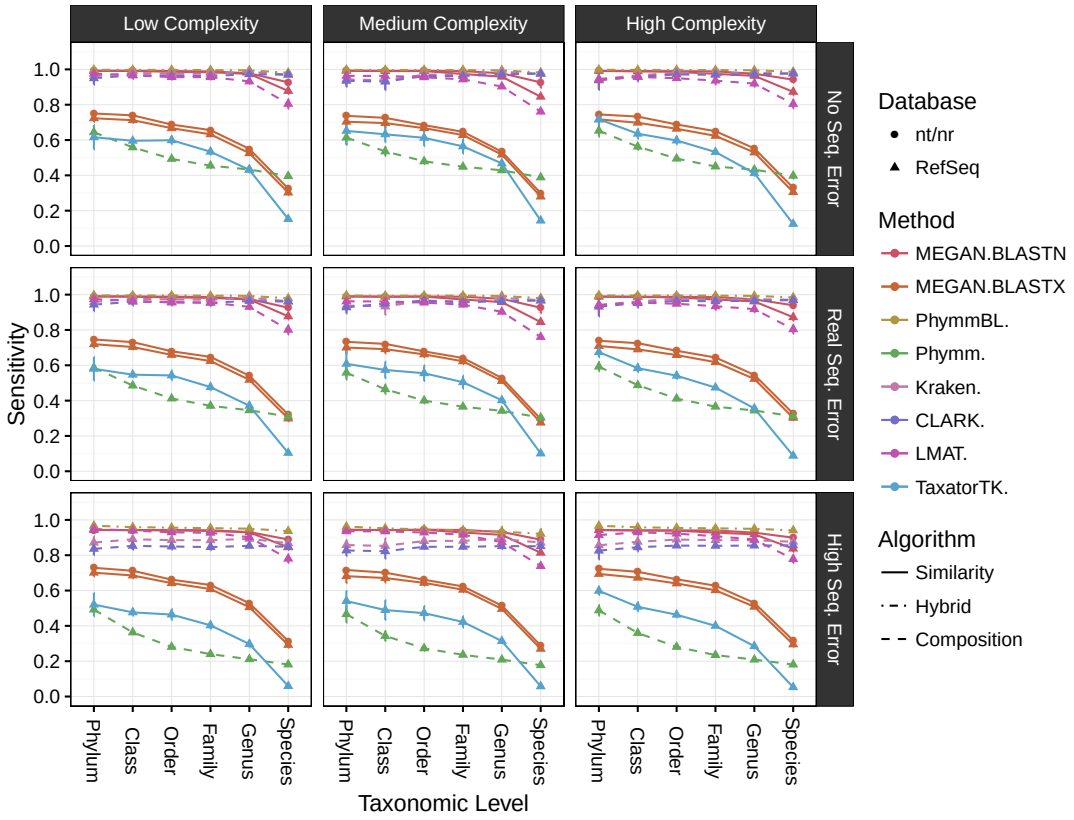


FIGURE 2.2 Sensitivity of taxonomic classification programs. Sensitivity of taxonomic classification programs was measured in triplicate at varying levels of community complexity and sequencing error. Sensitivity was examined at six taxonomic levels. The database used as a reference, the method, and the underlying algorithm governing each method are indicated. Error bars indicate the standard deviation of the mean.

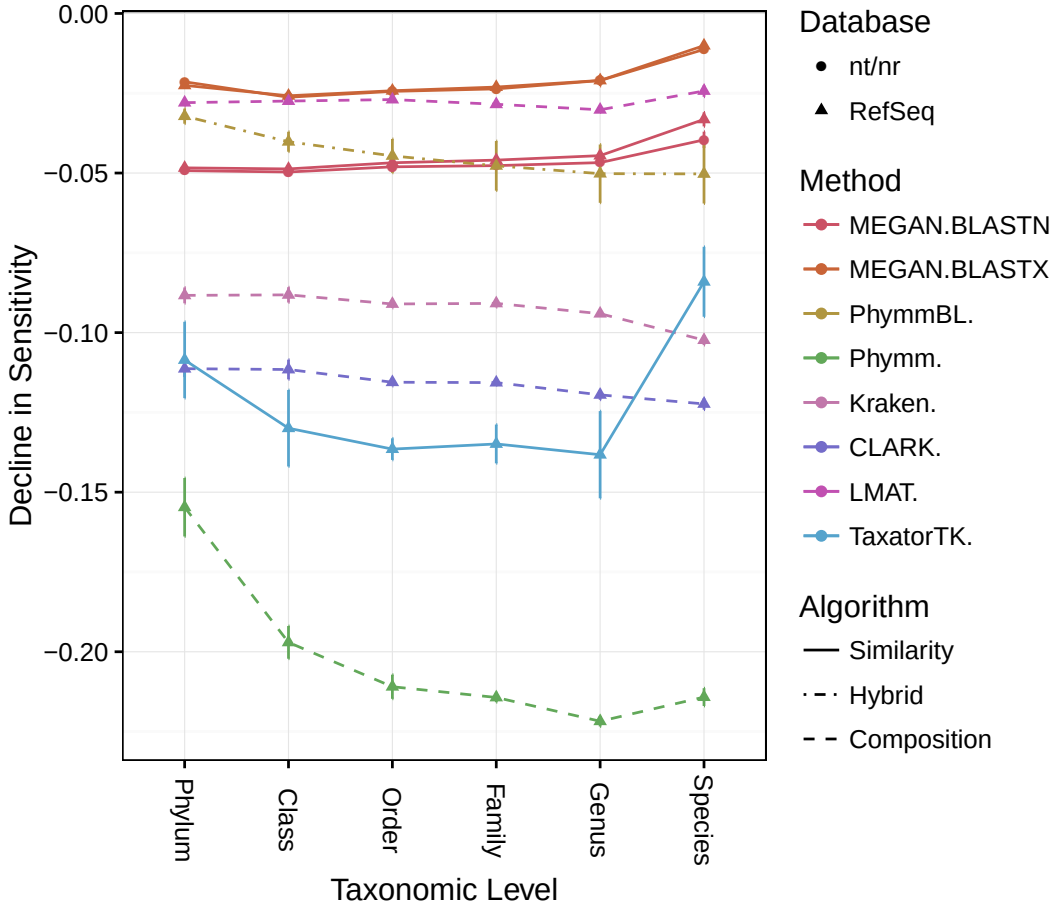


FIGURE 2.3 Sensitivity declines in response to sequencing error. The decline in sensitivity is defined as the difference between sensitivity when no sequencing error and high rates of sequencing error are present. The sensitivity across all community complexities was averaged to arrive at a mean sensitivity for each rate of sequencing error. A larger magnitude indicates a greater decline in sensitivity. All measurements are in triplicate and error bars indicate the standard deviation of the mean.

Testing classification efficiency when sequence reads are diverged from reference sequence using "evolved" sequences revealed a decline in specificity for PhymmBL only at less specific taxonomic ranks. Additionally, marked drops in sensitivity were observed for most programs, although the compositional programs CLARK and KRAKEN and similatiry-based MEGAN using blastn did not show a rank-dependent drop in sensitivity (Figure 2.5). PhymmBL showed the highest sensitivity at the species level when non-reference sequences were used, followed by MEGAN using blastn and the nt database. A slight decrease in sensitivity was seen as the species level for MEGAN using blastn with the RefSeq database, highlighting the potential effect of using a reduced database size in accurate detection at the species level. LMAT, CLARK,

and Kraken had similar sensitivity at the species level, although LMAT exhibited better performance at the genus or higher levels. MEGAN using blastx had comparable sensitivity up to the order level with MEGAN using blastn, but sensitivity declined strongly towards the species level.

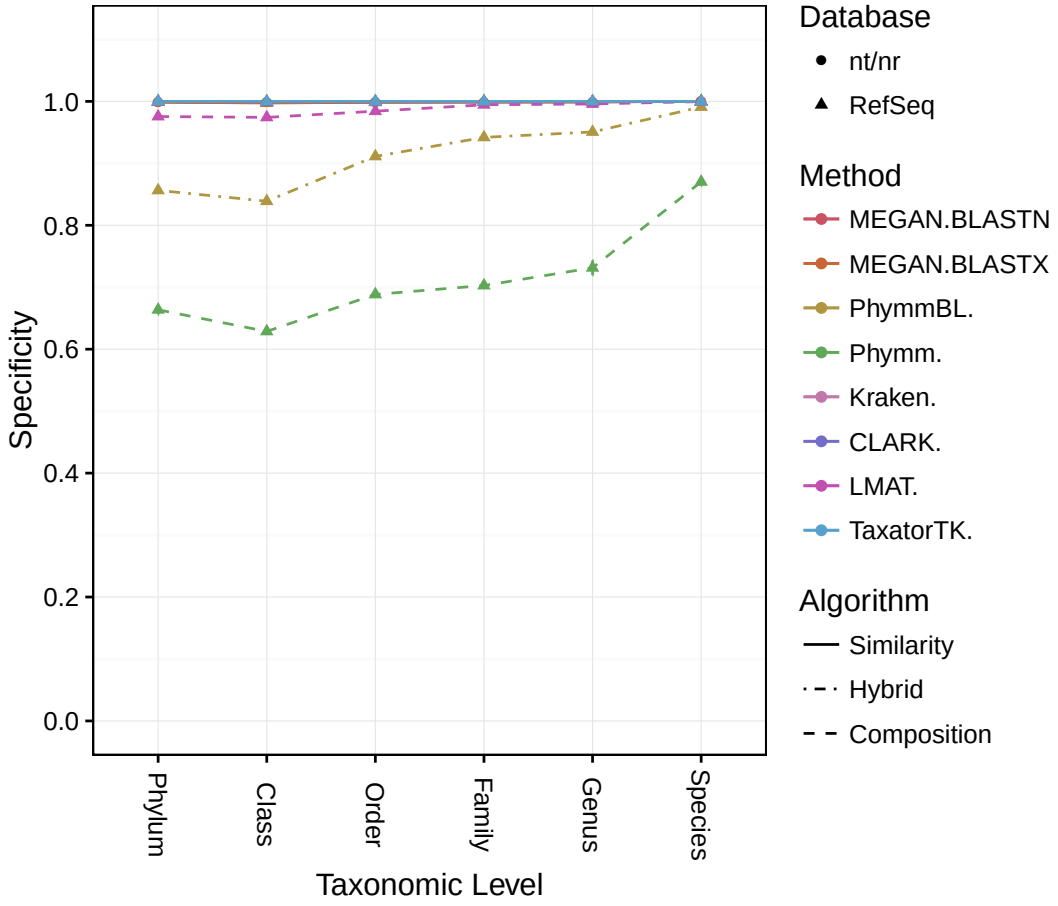


FIGURE 2.4 Specificity of taxonomic classification programs when reads are diverged from the reference sequences. Reads were obtained from genomes which were diverged from the reference genome present in the databases and specificity of taxonomic classification was measured at six taxonomic levels in triplicate. The database used as a reference, the method, and the underlying algorithm governing each method are indicated.

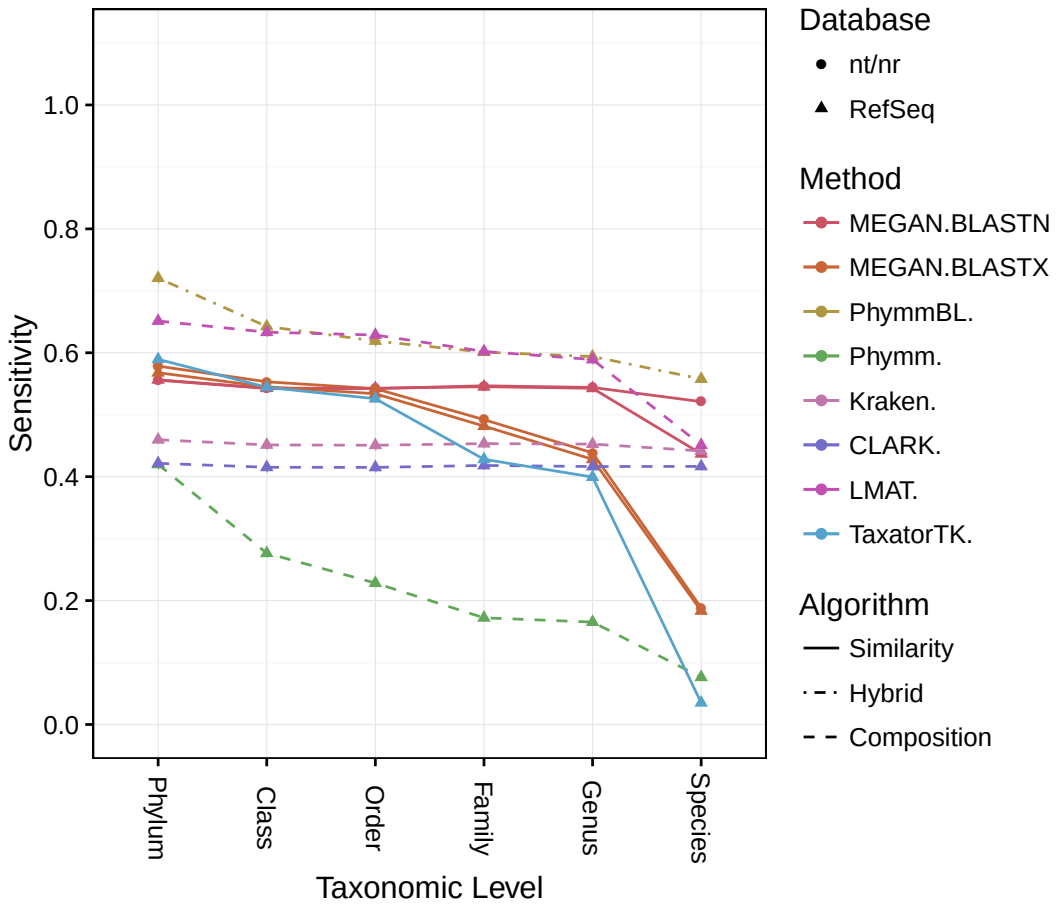


FIGURE 2.5 Sensitivity of taxonomic classification programs when reads are diverged from the reference sequences. Reads were obtained from genomes which were diverged from the reference genome present in the databases and sensitivity of taxonomic classification was measured at six taxonomic levels in triplicate. The database used as a reference, the method, and the underlying algorithm governing each method are indicated.

2.4 Discussion

The effects of different algorithms for taxonomic classification were explored by simulating metagenomes with reads of known origin. The utility of specificity as a metric did not seem to distinguish programs well, as most programs were correct when making a classification. In an introduction to the long fragment taxonomic classification program TACOA, Diaz et al. [68] found that specificity seemed to differ when analyzing varying fragment sizes, but even then variability was small. This finding, in combination with the findings of our analysis, suggests that specificity might be more

affected by analysis of taxonomic classification on different read sizes, but not between different programs at the same read lengths. The discrepancy between programs was more evident when sensitivity was examined. This indicated that there was varying success between the programs in capturing all the reads belonging to each clade. The variation in sensitivity also seemed to vary depending on the taxonomic rank and on the sequencing error introduced. Certain programs such as MEGAN using blastn and Taxator-tk showed a consistent drop in sensitivity as taxonomic ranks became more specific, while others like LMAT showed a sensitivity decline only at the species level. Sensitivity also declined as sequencing error increased, although MEGAN with blastx and LMAT did not seem to have a significant decline in sensitivity due to sequencing error. For blastx-based analysis, this is likely due to the redundant nature of codons and the introduction of synonymous mutations which do not alter the translated sequence. For LMAT, this could be due to the use of multiple overlapping k-mers and using an aggregated score to assign taxonomy. Since it is unlikely for most of the k-mers in a read to have mutations, and k-mers are assigned to an LCA, the "correct" k-mers will outnumber the k-mers with errors and a greater weight will be given to the correct taxonomic lineage. This also means that sensitivity at the species level might be impacted since classification will rely on k-mers associated with an LCA, and indeed, LMAT does exhibit a drop in sensitivity at the species level. Overall, for instances where reads contain reference sequences in the database, programs such as LMAT, CLARK, Kraken, PhymmBL, and MEGAN with blastn showed high sensitivity and would likely perform well in real analyses.

The effect of the database used as a reference was briefly explored by assessing MEGAN and blastn/blastx using the nt/nr or RefSeq databases. The RefSeq database is a subset of the nt/nr databases, yet for most purposes, it was sufficient in capturing the taxonomic composition accurately. The only cases where it seemed to have an effect was when the nt database was used with blastn at the species level. In this case, the nt database performed better in terms of sensitivity, especially when non-reference reads were tested. For blastx, using the nr database did not have any discernible effect. This suggests that the computational performance boosts associated with implementing a smaller database with the RefSeq database may outweigh the slight benefit in sensitivity when using the nt database for blastn analysis.

The lack of a clear trend between composition-based and similarity-based methods suggests that specific algorithms are more likely to dictate performance rather than the general type of methodology used. From the composition-based programs, LMAT performed best, while MEGAN using blastn and the nt database had the overall highest performance out of the similarity-based programs. PhymmBL, the hybrid program, had the greatest sensitivity in almost all scenarios at all taxonomic ranks, and performed the best when non-reference sequences were used. This suggests that hybrid methods may be more robust to divergence of sequences from the reference, a common occurrence in metagenomic studies. With the exception of the time-constraints imposed by this hybrid method, it should be considered for such studies, and more time-sensible hybrid methods should be developed.

The use of non-reference sequences is a more realistic representation of metagenomic studies since often times, reads do not originate from reference genomes. In this case, some studies implement the use of clade-exclusion as a means to test performance classification of non-reference reads [69], while others may use random shuffling of genomes to mimic the presence of unknown organisms in the dataset [70]. Clade exclusion works on the basis of removing the reference sequence of all reads in a dataset from the database. However, this might be biased towards several methods which implement the use of a LCA algorithm. In the case where two closely related species are in a metagenomic dataset, clade exclusion of both of those sequences from the database limits the information available for classifying each species independently. By removing all clades which are in the simulated dataset from the database, much more information from the database is removed than is needed to assess whether each species can be classified if its reference sequence is not present. For LCA-based methods where the use of closely related species is required to classify reads, this may reduce the performance drastically. One way to bypass this drawback is to only select species for simulation in the metagenomic dataset that are distantly related from one another so that the LCA algorithm is not affected by the removal of all the species, or to perform clade exclusion one at a time for each species. The former is hard to implement as this is typically not representative of real metagenomic datasets since many organisms inhabiting a given environment are typically closely related, and the latter is likely not feasible for a large number of species as a separate database would be required for each iterance of a clade exclusion (unless the program provides an option to temporarily remove a reference sequence, but almost all the programs tested do not have this option). Therefore, I opted to use a different approach by using sequences which are diverged enough from the reference sequences while maintaining the reference sequences in the database. This approach allows one to assess performance of non-reference sequences while not affecting the database, which is more representative of real world scenarios.

Using this method to assess performance of non-reference sequences, it was clear that this resulted in marked drops in sensitivity for all programs. It also revealed that all of the similarity-based programs performed equally well up to the order level, after which performance varied. Composition-based methods seem to have consistent performance at all taxonomic ranks, although sensitivity varied between programs and in general was greatest for LMAT and PhymmBL, although at the species level, MEGAN using blastn was also relatively highly sensitive.

Recent analysis conducted by Peabody et al. [69] and Lindgreen, Adair, and Gardner [70] set out to evaluate whole genome shotgun metagenomic methods for taxonomic classification. In the work conducted by Peabody et al. [69], the use of both *in silico* and *in vitro* metagenomic datasets was employed to assess performance. Only 11 species were included in their *in silico* analysis and only 12 species were assessed in the *in vitro* analysis. This is a major drawback of their assessment since this is not a realistic number of taxa that are typically present in a metagenomic dataset. The inclusion of such small numbers of taxa may affect the performance metrics obtained and may show high variability between replicates. This is mentioned in their paper and they

concede that "abnormal results from individual genomes could have a large impact on the results". In their analysis, replicates were not mentioned, and the small number of species highlights some flaws in their experiments.

Their assessment of performance was based on sensitivity and precision metrics which differed slightly from the metrics used in this work. Additionally, the programs selected for analysis differed and they utilized 250 bp sequences, so direct comparisons of results is slightly difficult, but general trends may be explored. Their findings suggest that CLARK and Kraken have poor sensitivity and precision, although they also note that very few reads were actually assigned for these two programs. This finding is likely due to the use of clade exclusion, as work by Lindgreen, Adair, and Gardner [70], which did not utilize clade exclusion, contrarily found CLARK and Kraken to have very strong performance. However, the use of *in vitro* analysis by Peabody et al. [69] did support their *in silico* findings. Peabody et al. [69] reported that MEGAN had relatively good performance, although they implemented RAPSearch2 rather than blastx and found that this outperformed MEGAN with blastn. This is in contrast to my findings where blastn-based MEGAN performed better. It is unclear whether the use of RAPSearch2 instead of blastx may be sufficient to increase performance sufficiently, but it is a possibility. The work by Lindgreen, Adair, and Gardner [70] and Peabody et al. [69] do not seem to be in agreement, however, the analysis by Lindgreen, Adair, and Gardner [70] agrees most with the analysis presented in this work. CLARK, Kraken, and LMAT performed well for taxonomic classification, while Taxator-tk was found in both analyses to be inadequate in terms of sensitivity, especially at more specific taxonomic levels. MEGAN using blastx performed worse than CLARK and Kraken in both studies, but performed better than many other options as well. Neither study from Peabody et al. [69] or Lindgreen, Adair, and Gardner [70] assessed the performance of PhymmBL, which was found to be the most consistently high performing taxonomic classification program in this analysis.

Finally, just as Lindgreen, Adair, and Gardner [70] allude to, although some programs outperformed others, the program of choice for metagenomic studies is dependent on the needs of the researchers. The computational requirements, run time, presence of a graphical user interface (GUI), rank-flexible assignments, functional assessment, and data visualization tools are all factors which can weigh on the decision of a taxonomic classification program. Although some programs performed poorly and would not be recommended for metagenomic analysis, the remainder of the programs may be each be utilized usefully depending on the context of the study and the needs of the researchers.

Chapter 3

Microbial Characterization of Freshwater Beaches

3.1 Introduction

Water quality monitoring of relevant aquatic environments is an important facet of public health. Areas of fresh and marine waters which receive regular exposure from the public are often monitored for their water quality to ensure the safety of those who access these areas. Typical water quality monitoring programs across the world are established with the intent of obtaining measures of chemically and biologically relevant indicators of water quality [71, 72]. In this chapter, water quality monitoring in the context of biologically relevant metrics and the application of metagenomics for this purpose will be focused on.

Waterborne pathogens pose a significant human health risk due to the potential for exposure of a large number of individuals. Exposure to and bathing in polluted coastal waters is estimated to result in over 120 million incidences of gastrointestinal illness globally [73]. Despite attempts by developed and developing nations alike to control outbreaks and emergence of water-borne pathogens through pollution sources, pathogens such as *Vibrio cholerae*, enteropathogenic *Escherichia coli*, and *Campylobacter* spp. regularly persist in water bodies and sometimes result in waterborne outbreaks [74]. The goal of monitoring water quality to detect such pathogens and controlling public exposure to potential illness is challenged by the inability to effectively detect these pathogens in a timely manner. Culture-dependent methods are often not capable of recreating the delicate growth conditions of many pathogens [10, 16], and in cases where it is possible, it may take many days to give rise to a quantifiable measure of the pathogen load (e.g. culturing of *Campylobacter* spp. and many *Legionella* species can take up to 5 days [75, 76]). Within the context of public health, the lag between the upsurge of pathogens in a water body and the detection, quantification, and reporting of that event must be minimized for effective health risk aversion.

Cases of bathing-associated gastrointestinal illness are often caused by a diverse number of fecal pathogens that are introduced into the aquatic environments through

wastewater treatment plants, sewer overflows into water bodies, urban and agricultural runoff, and from local human and animal populations [77]. Species which serve as indicators for these sources of pollution (e.g. fecal indicating bacteria (FIB)) can provide a predictive capability for the potential of illness in beach goers. The use of indicator species as an indirect measure of the pathogenic load in a water body has therefore been a staple in water quality monitoring programs and has been shown to correlate well with water quality and gastrointestinal illness in recreational waters [78]. Although culture-dependent methods are the predominant form used by water quality monitoring programs [79, 80], chromogenic substrate [81] and quantitative PCR (qPCR) methods [78] are some alternate methods which have been utilized and also validate the correlation of indicator bacteria and incidence of illness in swimmers of recreational waters. Various indicator species have been established, and the incidence of illness in recreational waters correlates strongly with indicators of human and animal fecal contamination such as *Escherichia coli*, *Enterococcus*, and fecal and total coliform counts. However, these indicators typically show heterogeneity in their predictive capacity between different water environments or sources of contamination. For instance, while *Enterococcus* is an effective indicator in marine waters, its capacity as an indicator in freshwater environments is not as effective, highlighting the environmental dependency of *Enterococcus* as a water quality indicator [82]. In freshwater and inland beaches where point sources of contamination are established, *E. coli* and fecal and total coliforms have been shown to be a strong indicator for gastrointestinal illness in swimmers [80, 82]. However, in areas where non-point sources of fecal contamination are predominant, these same indicators were shown to not be associated with health risks [81].

Despite the heterogeneity in the capacity of various indicator bacteria to predict health risk and illness, there is widespread use of these bacteria in water quality monitoring programs due to the lack of a more comprehensive system for the detection of pathogens. Although culture-independent methods for pathogen quantification such as qPCR may circumvent the challenges of culturing, they are limited in their scope of the number of pathogens that can be monitored simultaneously. Bypassing the use of indicator organisms and directly probing environments for pathogenic bacteria as they exist in the complete bacterial population would alleviate many of the challenges and drawbacks that are associated with an indicator or single species based method for water quality monitoring [83]. The use of DNA microarray technology has been previously proposed, but no developments into this method have taken place for use in water quality monitoring [83]. A metagenomic approach which can directly detect numerous waterborne pathogens would therefore prove useful since a link between indicators and pathogens is no longer required, and investigation into the complete microbial community and all pathogens would be possible. This could provide a more comprehensive assessment of the water quality and improve water quality monitoring programs and the decisions they make.

An additional aspect of water quality monitoring that is often overlooked is the presence of potential pathogenic microorganisms inhabiting sand environments of recreational beaches. Previous studies have consistently shown the presence and persistence of *E. coli* and *Enterococcus* in wave-washed sands, giving rise to the potential of microbial pathogens in these environments [84–86]. Using culture-based methods, *E. coli* and *Enterococcus* were found to be abundant in wet sands of freshwater beaches, and showed up to 35 greater concentrations in the sand, suggesting that the sand at these beaches could act as a reservoir for these organisms [87]. As well, source tracking of *E. coli* in beach sands revealed that some of the *E. coli* strains were autochthonous and were transmitted to the water column [88]. Additionally, the detection of *Aeromonas* spp. [89], *Campylobacter* spp. [90, 91], *Salmonella* spp. [90, 91], and *Vibrio* spp. [92] in various sand types of marine and freshwater beaches suggests that this environment is inhabitated by numerous pathogens and has important public health impacts. Despite the notable exposure of the public to these pathogenic and pathogen indicating microorganisms in beach sands, most monitoring programs do not survey these environments. If there are distinct microbial differences between the sand and water environments at the same beach, then solely monitoring the water may not be sufficient for public health.

To further expand upon the findings regarding beach sands and indicator organisms, as well as establish a more comprehensive methodology for the monitoring of water quality, we employed a metagenomic approach to characterize microbial populations in beach sands and waters. Since indicator microorganisms may not be consistent in every environment and are not a direct measure of the potential pathogenic load in an environment, a metagenomic approach that can directly probe for pathogens would be extremely beneficial. Additionally, although differences in *E. coli* and *Enterococcus* exist between beach environments, a more comprehensive comparative analysis of freshwater beach sand and water environments has not been conducted. In the interest of public health, it is imperative that a comparative analysis of these microbial communities is examined.

To address the issues which persist in the current standards for water quality monitoring, and to build upon the sparse understanding of beach sands, a comprehensive analysis of microbial communities of four freshwater beaches in the Niagara Region was investigated. The Niagara region in Ontario, Canada is home to over 25 public freshwater beaches open for recreational use across two of the five Great Lakes [93]. As a popular attraction for many residents of the area, water quality of these beaches is regularly monitored using coliform counts of *E. coli* [93]. This region serves as a suitable testbed for the development and potential implementation of a metagenomic approach for monitoring of pathogens, and also allows for the investigation of important factors which influence microbial populations.

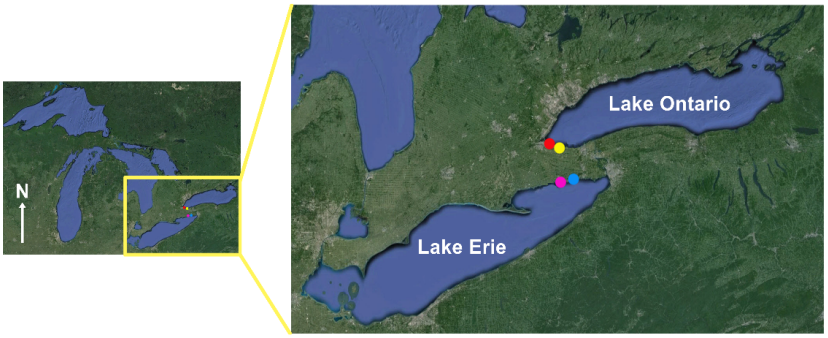
3.2 Methods

3.2.1 Sampling

Four beaches were selected for sampling based on their importance in the Niagara region watershed and geographic dispersion across two of the five Great Lakes; Lake Erie and Lake Ontario. Long Beach and Nickel Beach of Lake Erie and Lakeside Beach and Fifty Point Beach of Lake Ontario were sampled during the summer months of 2012 and 2013 and 16 pairs of samples were collected across all four locations (Table 3.1, Figure 3.1a). Each pair was comprised of a sample corresponding to the beach water and beach sand habitat. Water samples were collected in sterile $1.0\,\mathrm{L}$ sampling bottles (Nalgene) at the surface level of $1\,\mathrm{m}$-deep water. Sand habitats were sampled through the generation of $30\,\mathrm{cm}^3$ sand pores $1\,\mathrm{m}$ into the dry supratidal zone from the intertidal zone and subsequent collection of the water which permeated through the sand into the pore (Figure 3.1b). Sample collection was followed by storage on ice and processing for extraction of DNA within 6 hours. $300\,\mathrm{ml}$ of the water collected from each sample was filtered through a $0.45\,\mu\mathrm{m}$ membrane (Fisher Scientific) and the resulting retentate was extracted as the bacterial fraction. DNA was extracted from the bacterial fraction through the use of the PowerSoil DNA Isolation Kit (Mo Bio Laboratories, CA, USA) according to the suppliers standard extraction procedure.

TABLE 3.1 Sampling Sites

Beach	Lake	Years Sampled	Samples
Long Beach	Erie	2012	8
		2013	10
Nickel Beach	Erie	2012	2
Lakeside	Ontario	2012	6
		2013	4
Fifty Point	Ontario	2012	2



(A) Sampling Locations

(B) Sand Pore Sample

FIGURE 3.1 Sampling sites and procedures. 16 pairs of samples were collected in total from four beaches throughout 2012 and 2013: Lakeside Beach (yellow) and Fifty Point Beach (red) of Lake Ontario and Long Beach (pink) and Nickel Beach (blue) from Lake Erie. Each pair of samples corresponds to one water sample and one sand pore sample as seen in 3.1b.

3.2.2 Library Preparation and Sequencing

DNA extracted from the bacterial fractions was subjected to shotgun metagenomic sequencing. DNA was diluted to $0.2\,\mathrm{ng}\,\mu\mathrm{l}^{-1}$ and prepared using the Nextera XT DNA Library Preparation Kit (Illumina). Samples were sent to the Farncombe Metagenomics Facility (McMaster University, Hamilton, Ontario, Canada) for 100 bp or 150 bp paired-end sequencing on the Illumina HiSeq 2000 platform.

3.2.3 Bioinformatic Processing

Sequence reads obtained from Illumina HiSeq 2000 sequencing were processed for low quality bases and adapter sequences using BBDuk (version 35.82) and quality-screened using FastQC (version 0.11.3) (Figure 2). BBDuk processing parameters were set to

right- and left-end quality trimming with right side adapter trimming using a kmer size of 12 with the Nextera adapter sequences. Zhou et al. defined high quality bases as having a quality score greater than 20, and this was therefore selected as the threshold [94]. Default parameters implemented in the web-based metagenomic analysis software MG-Rast set the quality threshold at 15, further justifying the use of 20 as the threshold in BBDuk. A minimum length of 40 after trimming was also imposed. Reads were trimmed in a paired-end fashion, however, singleton reads whose other pair was discarded were retained for downstream analysis. All samples passed FastQC analysis after BBDuk processing and were retained for further analysis. Processed reads were compared against the NCBI non-redundant protein database using DIAMOND blastx (version 0.7.8.57) [66]. Reads obtained from BLAST comparison were parsed and classification was assigned using MEGAN (version 5.10) [54] with default classification parameters (minimum bit-score 50; max e-value 1e-3; top percent 10; min support 5). MEGAN was used to perform classification for both taxonomic and functional annotation, which utilizes the NCBI taxonomy [95] and SEED subsystems [96] databases for assignment. Each member of a paired read was classified independently, and a custom Perl script was used to arrive at a consensus classification for a read pair using the average bit score of the top 10% BLAST hits from each member of the pair.

For pathogen detection, reads were taxonomically assigned at the species level to the bacterial RefSeq database using CLARK (v1.1.2) [63]. Paired reads were classified jointly and singletons were appended to the classifications. A minimum count of 5 was imposed for detection and a confidence score determined by CLARK of 95% was used as a threshold for classification.

3.2.4 Statistical Analysis

All statistical analysis was conducted in `R` (version 3.2.3). Ecological diversity analysis was accomplished with the `vegan` package (version 2.3-4). `DESeq2` (version 1.10.1) was used for statistical testing of differential abundance of taxonomic and functional features. The model for statistical testing was implemented with sample pair information, accounting for the pairing of the data.

More detailed information about specific statistical methods is included within the results text and corresponding sections.

3.3 Results

3.3.1 Sequence Metrics

A major goal of this work is to compare the microbial communities inhabiting beach waters and sands. It is therefore imperative that sampling, sequencing, and bioinformatic processing do not impose any form of bias on one of the two environments

asymmetrically, which would otherwise hinder the interpretation of the downstream results. To ensure that no such bias exists, the number and quality of reads present in each sample was examined (Figure 3.2).

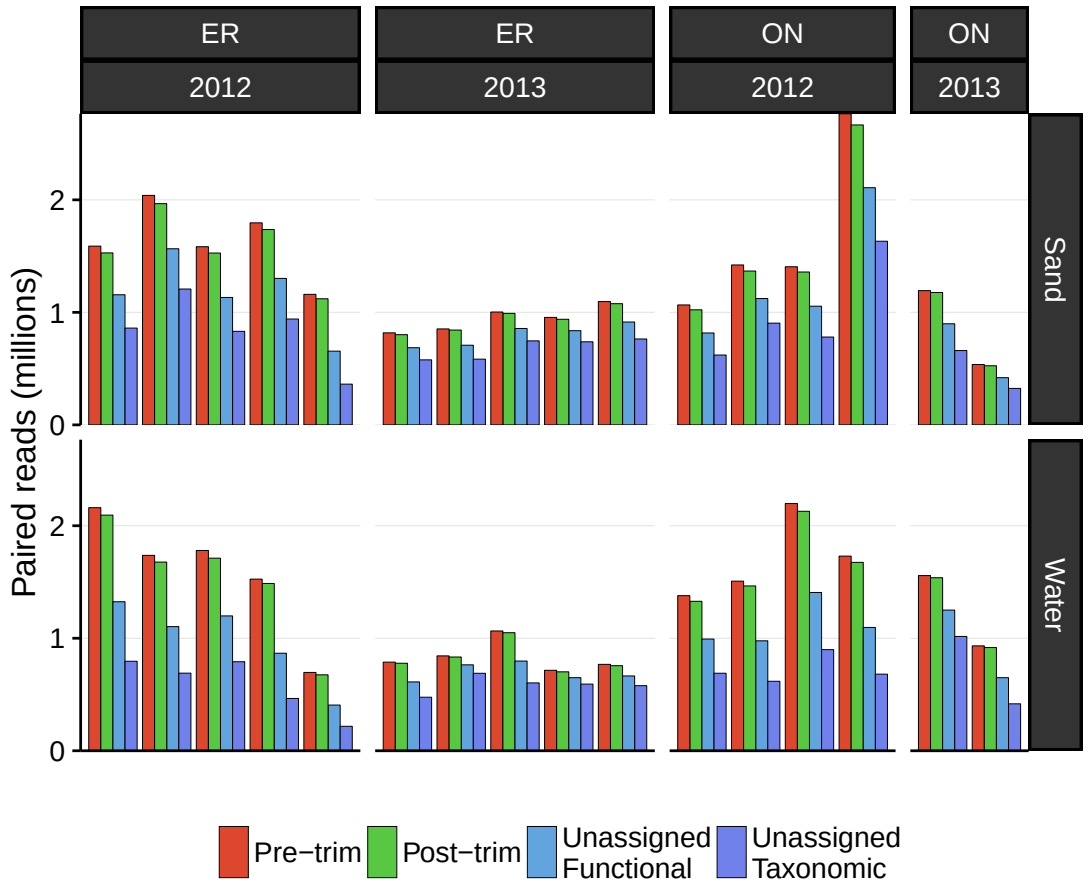


FIGURE 3.2 Read processing statistics. All values are represented as millions of paired reads. Post-trim values correspond to number of pairs in addition to singleton reads whose partner did not survive quality control. Unassigned functional and unassigned taxonomic values represent the number of post-trim reads minus the number of reads with a functional or taxonomic assignment, respectively. The samples are grouped by lake and year and separated by sample type.

On average, there were 1,333,106 paired reads before processing and 1,295,650 paired reads after processing across all samples. This is an average loss of approximately 8.7% of reads due to processing, indicative of relatively high quality reads overall. After quailty and adapter trimming, average read qualities were above 32 for all samples with standard deviations not exceeding 2.5. Between sample types (pore or water), there was no significant difference in the number of reads prior to or after

processing, suggesting that there was no bias in the sample collection or processing between the two sampling environments (Table 3.2; t-test, $p > 0.5$). There were significantly more reads in the 2012 processed samples than the 2013 processed samples, however, this is attributed to a larger lane capacity per sample in the 2012 samples (t-test, $p = 8.67 \times 10^{-6}$, see metadata in B.3). Since pore and water samples are mixed well between the two years due to the paired nature of the samples, the effect on the comparison of the sand and water habitats is minimized. The proportions of reads unassigned to a function or taxa were not different between environments (Table 3.2), suggesting that there was no bias for classification during the bioinformatic processing in either environment. In terms of read metrics, there does not seem to be a bias in quality or classification as a result of the sampling environment, an important assessment which allows for more conclusive and accurate comparisons between the two environments using this data.

TABLE 3.2 Read Metrics by Sample Environment

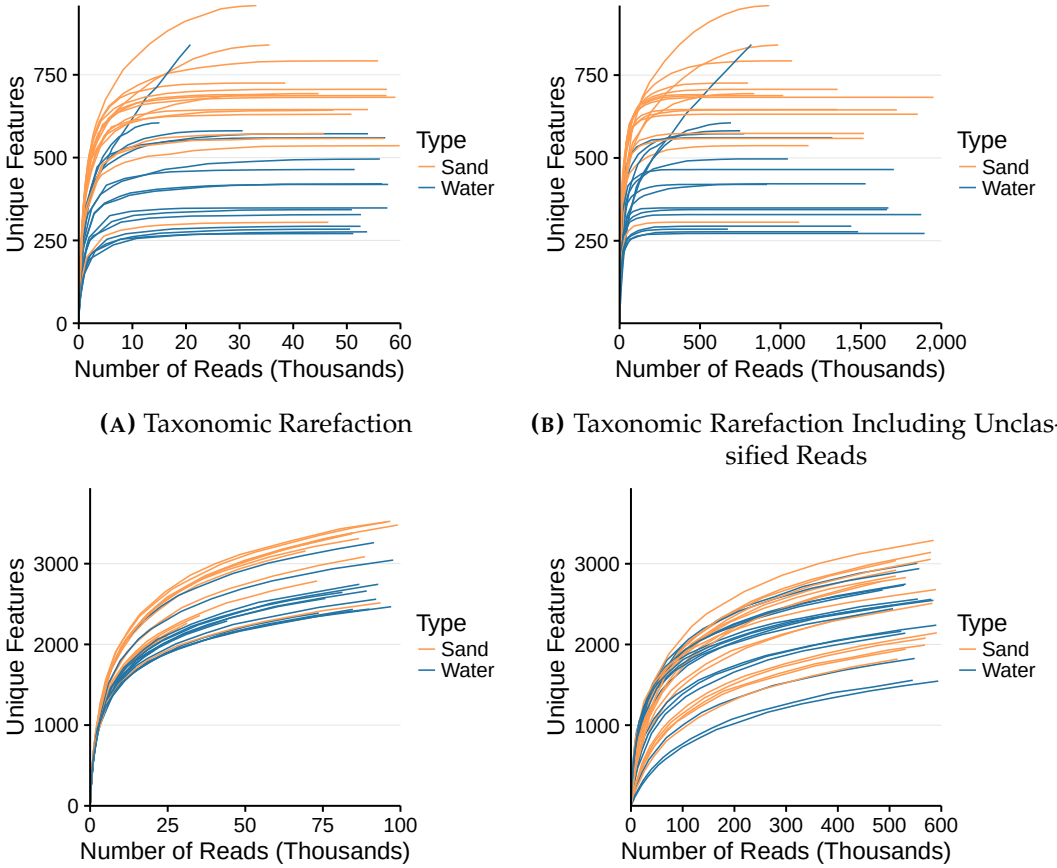
Metric	Sand	Water	Average
Paired Reads (pre-trim)	1,329,907	1,336,305	1,333,106
Paired Reads (post-trim)	1,290,357	1,300,942	1,295,650
Proportion Unassigned Functions	0.76	0.76	0.76
Proportion Unassigned Taxa	0.58	0.56	0.57

3.3.2 Richness and Diversity

Rarefaction curves are an established method for the determination of sampling saturation and the effect of sampling depth on feature (taxonomic or functional) discovery [97]. Determination of sampling saturation is important when attempting to make inferences on differential presence or absence of taxonomic or functional features, and also provides information about the nature of the sample and the adequacy of the sequencing depth. Samples whose rarefaction curves do not plateau indicate that the sequencing depth was not enough as greater subsets of the data continuously reveal new features. Rarefaction curves which do plateau suggest that sequencing depth was adequate to capture the community composition as greater subsets of data do not reveal new features. A disparity in saturation between conditions of a parameter of interest may suggest that sequencing depth requirements differ between those conditions, and inferences about presence or absence of features may become unclear since the lack of presence of a feature in the data no longer necessarily corresponds with lack of presence in actuality. It is therefore important to establish the levels of sampling saturation through the use of rarefaction curves prior to downstream analysis, especially within the context of the comparative analysis of microbial communities between beach sands and waters.

Rarefaction curves are often applied to amplicon sequencing data (such as 16s rDNA amplicon sequencing) since the number of unique 16s sequences directly correlates with the number of unique taxonomic features. However, rarefaction curves applied to shotgun metagenomic data can still be informative despite the loss of equivalence of unique sequences and unique features. A caveat of rarefaction of shotgun metagenomic data is whether unclassified reads should be included in the subsetting process. Excluding unclassified reads from the rarefaction curve means that subsetting only occurs on classified reads. This would not be an accurate measure of sequencing depth, but would still be informative about the nature of the feature discovery (i.e. increasing or plateauing). Including unclassified reads in the subsetting process would provide information on sequencing depth but would not be representative of the true unique features since the unclassified reads could belong to as many unique features as there are unclassified reads. This is due to the inability to distinguish unique features from one another when reads are unclassified and can potentially originate from any part of the genome. It is therefore important to assess how the inclusion of unclassified reads affects the rarefaction patterns, if at all. Despite this, rarefaction curves did not differ in the overall trends when excluding or including unclassified reads (Figure 3.3). Rarefaction curves show similar trends in either case, and only the scale of the x-axis on which the rarefaction is displayed differs.

Taxonomic rarefaction reveals that sampling is saturated in terms of discovery of unique taxonomic features and the saturation is not biased between beach environments (Figures 3.3a, 3.3b). It is also interesting to note that for the majority of samples, plateauing occurs at a relatively small number of reads, suggesting that detection of the same taxonomic features could be accomplished with a smaller sequencing effort. However, it is important to keep in mind that shotgun metagenomics is utilized for its capacity to yield information about functional capacity. A smaller sequencing effort may yield a similar capture of the taxonomic composition, but as seen in the functional rarefaction curves (Figures 3.3c, 3.3d), it would result in a weaker capture of the functional capacity of the communities since the functional rarefaction curves do not show saturation and unique features are detected as larger subsets of the sample are incorporated.



(A) Taxonomic Rarefaction

(B) Taxonomic Rarefaction Including Unclassified Reads

(C) Functional Rarefaction

(D) Functional Rarefaction Including Unclassified Reads

FIGURE 3.3 Taxonomic and functional rarefaction curves. Rarefaction curves were generated for taxonomic and functional features by subsetting reads and counting the number of unique features detected at each subset. All subsets were done with 10 replicates. The environment type is indicated by the colour of the lines. **(A)** and **(C)** do not include unclassified reads in the subsets, while **(B)** and **(D)** do.

Although most of the samples displayed taxonomic saturation, one water sample from Lake Erie did not (Figure 3.3a). This corresponds to a 2013 Long Beach water sample (Sample_72_BF17) and suggests that the taxonomic composition of this sample was not fully captured since a greater sequencing depth would lead to the detection of more unique taxonomic features. The corresponding paired sand sample did not show the same trend, and other water samples from the same beach in the same year also did not show this trend. This specific site did show a disproportionate deviation in richness from other water communities and had the highest richness among them, although diversity was not abnormal (see Figure B.4 in Appendix). The presence of numerous low abundance taxa resulting in elevated richness is likely resulting in the

rarefaction trend seen and is consistent with the lack of diversity increase. The removal of low abundance taxa (taxa with less than 10 reads) results in the rarefaction curve resembling that of the remaining samples, serving as further verification of the presence of many low abundance taxa.

Another important trend gathered from the rarefaction curves is that there seems to be a greater number of unique taxonomic features detected in the sand than in the water (Figure 3.3a). This is further validated by examining taxonomic richness between the sand and water habitats (Figure 3.4a). Unique taxonomic features were, on average, greater in the sand (667) than in the water (444) (paired t-test, $p = 3.40 \times 10^{-5}$). This was true for all pairs with the exception of the pair containing the site at which taxonomic saturation was not reached. On average, richness was greater in samples obtained in 2013 (655) than in 2012 (478) (t-test, $p = 5.66 \times 10^{-3}$). Although there was a significant difference in the total number of reads gathered between years, the presence of taxonomic saturation in samples from both years and the fact that the richness is inverse to what is expected by read counts alone does not suggest that the difference in taxonomic richness is a consequence of differences in sequencing depth but rather is a result of true taxonomic differences. Additionally, there was not a significant difference in richness between lakes (t-test, $p > 0.05$). To investigate a potential interaction effect of the sample type and year and to establish whether beach habitat differences are temporally-dependent, a linear model was fitted to the data. The linear model confirms the previous findings and further indicates the absence of an interaction effect of sample type and year as this term was not significant (Table 3.3). This finding suggests that there was not a temporal dependency of the increase in taxonomic richness in the sand and that the type effect was independent from year-to-year variations in richness.

TABLE 3.3 Analysis of Variance of Linear Model for Taxonomic Richness

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	1	396940.50	396940.50	25.48	**0.000027**
Year	1	247646.00	247646.00	15.90	**0.000458**
Lake	1	7844.93	7844.93	0.50	0.483987
Type:Year	1	8534.67	8534.67	0.55	0.465554
Residuals	27	420559.89	15576.29		

Since taxonomic classification was conducted with MEGAN's LCA algorithm, it is important to establish the effect of classifications which occur at a higher, less specific taxonomic rank. The significant differences in the number of unique taxonomic features between environments and years could be a consequence of disparity in the taxonomic classification specificity between these variables. For instance, if many of the classifications in the water samples occurred at a higher taxonomic rank, then the number of unique taxonomic features would be underestimated, leading to an artificially significant enrichment of richness in the sand. To determine the effect of the LCA algorithm on unique taxonomic feature detection, each taxonomic rank from 'Phylum' to 'Species'

was assigned a value from 1 to 6, respectively. Averaging these values for each sample across reads assigned to unique taxonomic features, referred to here as the taxonomic rank index, provides a measure of the average rank to which reads were assigned and can address the effect of the LCA algorithm. Between sample types, there was a significant difference in classification specificity (t-test, $p = 5.94 \times 10^{-7}$; Figure 3.4b), with the index being slightly higher in the water samples. This suggests that the greater richness seen in the sand habitat was not an artifact of the LCA algorithm and contrarily suggests that the number of unique taxonomic features in the sand may be underestimated in comparison to the water habitat since classifications were on average less specific. There was not a significant difference in classification specificity between years, again suggesting that the differences in taxonomic richness are not an artifact of the classification algorithm (t-test, $p > 0.05$; Figure 3.4b).

Across all samples, an average of 555 unique taxonomic features were detected. It is likely that the use of the LCA algorithm severely underestimates the number of unique taxonomic features since unique hits are relatively rare for blastx-based taxonomic analysis. Indeed, on average, only 10% of reads with a blast hit were unique hits, revealing a large number of reads which exhibited similarity with potentially multiple species. These reads would therefore be assigned to less specific taxonomic ranks and not be considered unique. For instance, if a read showed similarity to a species not yet detected in the sample, but also had hits to multiple other species, it may be assigned to the 'Family' classification common to all the species. If other reads have a classification under that same family, then this read will not be considered unique. This scenario is a drawback of the LCA algorithm is the likely cause for the relatively small number of unique features detected.

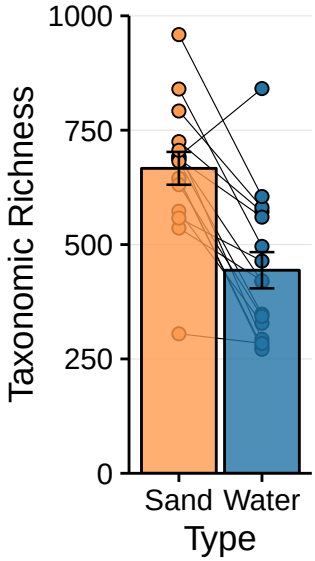
Functional richness was examined by determining the number of unique SEED-categorized functions detected within samples. Since functional saturation was not reached and the number of reads in samples collected in 2012 was greater than in 2013, it is expected that functional richness should be enriched in the 2012 samples. This expectation was verified with strong significance (t-test, $p = 3.41 \times 10^{-9}$; Figure 3.5a), highlighting the impact of differential read metrics on interpretations of ecologically relevant measures when saturation is not reached. Examination of functional richness between sample types reveals a slight enrichment in the sand than in water (paired t-test, $p = 6.27e - 03$; Figure 3.5a). This may suggest that the increase in taxonomic richness in sand habitats translates to functional richness and may indicate that unique taxa between the sand and water may harbor unique functions.

The greater taxonomic richness in the beach sands was also complemented with significantly greater species diversity as assessed by the Shannon index measure of alpha diversity (paired t-test, $p = 6.51 \times 10^{-4}$; Figure 3.4c). This pattern was consistent across both Lake Erie and Lake Ontario. Similarly, samples from 2013 exhibited significantly greater taxonomic diversity, in line with the richness analysis, and this pattern is again absent between lakes (Table 3.4). The alpha diversity complemented with the

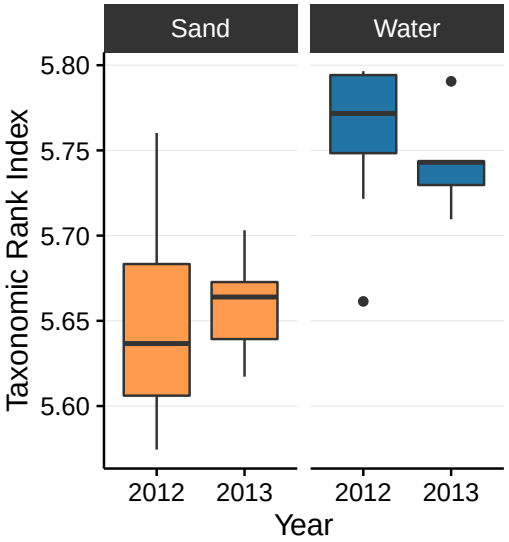
species richness analysis suggests that sample type and year are more important predictors of these measures than the lake where sampling was conducted. However, our analysis is limited to two years and two lakes, so these findings should be extended to a larger number of sampling locations and time points to further validate the conclusions. Functional diversity was also enriched in beach sands, in line with the richness analysis ($p = 1 \times 10^{-3}$; 3.5b). This effect also extended between years more significantly and is likely due to the discrepancy in sequencing depth between years (Figure 3.5b).

TABLE 3.4 Analysis of Variance of Linear Model for Taxonomic Diversity

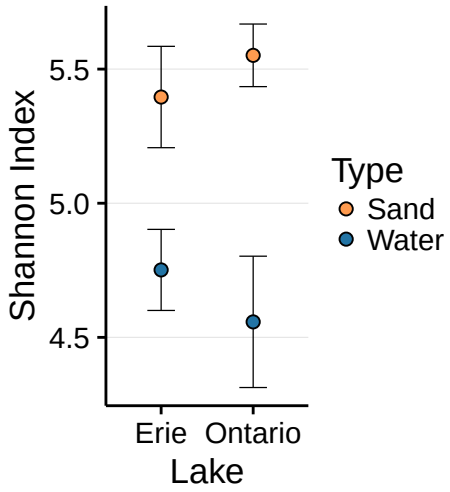
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	1	2.58	2.58	11.27	**0.0024**
Year	1	1.71	1.71	7.48	**0.0109**
Lake	1	0.01	0.01	0.04	0.8391
Type:Year	1	0.19	0.19	0.81	0.3763
Residuals	27	6.19	0.23		



(A) Taxonomic richness is greater in sand than in water. Taxonomic richness was measured as the number of unique taxonomic features detected. Richness for each sample is plotted as a point connected to the corresponding member of its pair. Error bars indicate the standard error of the mean.

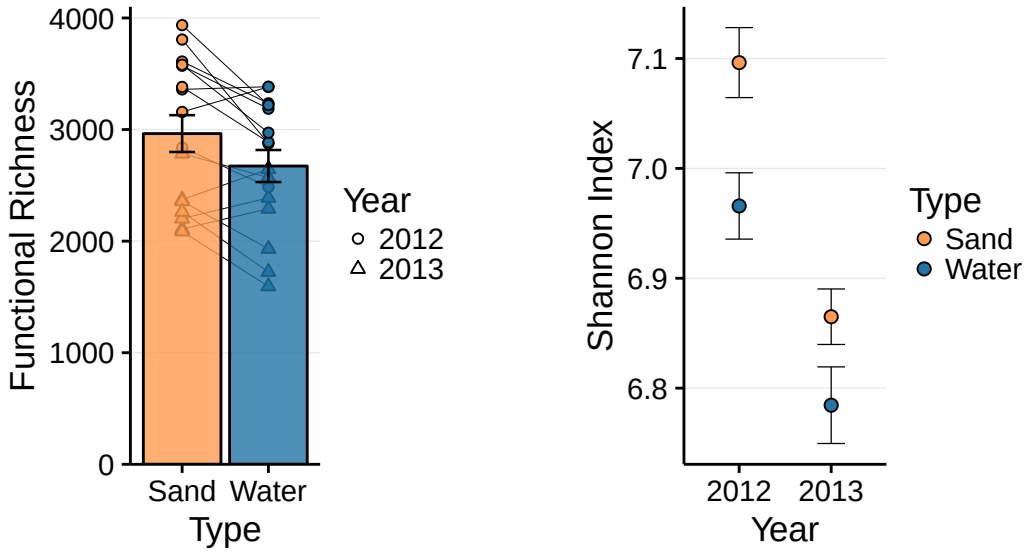


(B) Average rank of classifications is more specific in water than in sand. Boxplot of the ranks at which reads were classified across sample types and years. Larger values indicate more specific taxonomic classification (maximum of 6 for species level classification).



(C) Bacterial populations are more diverse in sand than in water. The Shannon Index was used to measure alpha diversity within each sample and averaged between environment types and lakes.

FIGURE 3.4 Richness and diversity quantification of bacterial populations in beach sands and waters of the Niagara Region.



(A) Functional richness is greater in sand than in water. Functional richness was measured as the number of unique SEED-categorized functions detected. Richness for each sample is plotted as a point connected to the corresponding member of its pair. Error bars indicate the standard error of the mean.

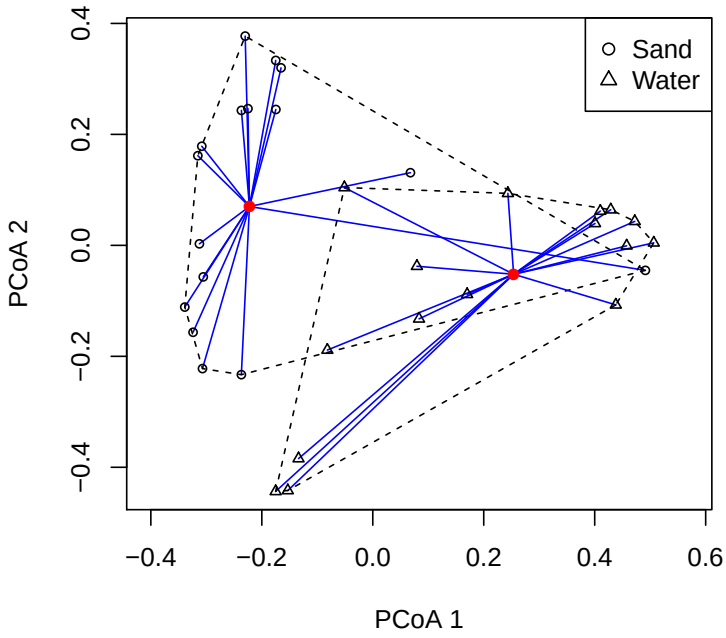
(B) Functional diversity is impacted by sample type and year of sampling. Functional diversity was measured using the Shannon Index within each sample and averaged between sample types and years.

FIGURE 3.5 Richness and diversity quantification of functional capacity in beach sands and waters of the Niagara Region.

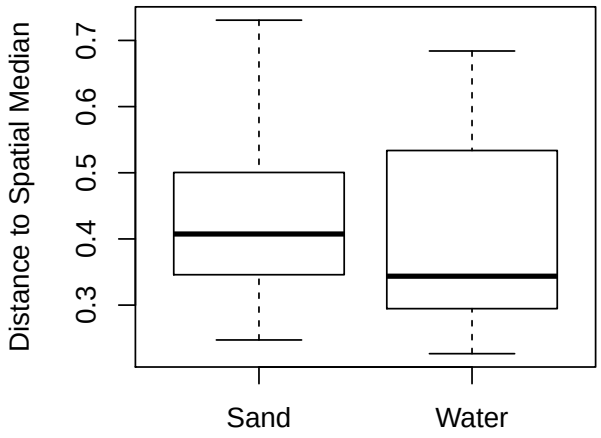
Measures of beta diversity are informative of regional differences of ecologically relevant sites [98]. In this case, beta diversity can be applied to studying diversity of sites within groups of interest, where the groups can be the sampling environment (sand or water) and sites can be each sample within those environments. Although several methods for the calculation of beta diversity have been established, most of these measures typically rely on presence/absence data and regard beta diversity as the ratio of the collective richness of a group of sites to the average richness of each site within that group, where groups can be defined by any variable of interest [98]. Since these indices rely on richness, there is an inherent loss of information because relative abundance of the taxonomic constituents is not considered. Beta diversity measures which utilize a dissimilarity index are informative of relative abundance of taxonomic constituents and attempt to address this issue [99]. Anderson, Ellingsen, and McArdle [99] have proposed a definition of beta diversity as the average distance, as determined by a dissimilarity measure, of every site (e.g. sample) within a group (e.g. environment) to the centroid of that group within multivariate space. Since a dissimilarity measure is used, the abundance of taxonomic constituents can be accounted for. This multivariate dispersion can then be used as a measure of beta diversity and allows for the ability to

statistically test differences in beta diversity between groups of interest.

The Bray-Curtis dissimilarity measure was used to determine all pairwise dissimilarities between samples on the basis of taxonomic constituents at the species level. Although several dissimilarity measures are used in ecological studies, this dissimilarity measure has been utilized in ecological contexts and metagenomic contexts with strong robustness and provides an intrinsic normalization between different sample sizes, making it a strong candidate for this analysis [98, 100]. The `betadisper` function from `vegan` was used to determine the multivariate dispersion between sample types. Multivariate dispersion was not significantly different between sand and water communities, suggesting that beta diversity and the collective diversity of these groups does not differ from one another ($Pr = 0.53$; Figure 3.6). One sand sample seemed to cluster strongly with the water samples and is closely associated with its corresponding water sample. This was true in principal coordinates analysis (PCoA) at all taxonomic levels as well (Figure B.6). This pair of samples is from Nickel Beach, and suggests that either the environment type at this beach does not dictate bacterial composition or the collection process did not accurately compartmentalize these two environments. Removal of these two samples does not affect the significance of the multivariate dispersion measure ($Pr = 0.93$).



(A) Beta diversity does not differ between beach environments. Beta diversity was measured by multivariate dispersions of sites to the spatial median in multivariate space. Spatial medians are indicated by red points and samples are indicated by hollow circles or triangles. The distances between each site to the spatial median is indicated by the blue lines, and the complete structure for each group is outlined by the dashed line. The average distance between sites and their corresponding spatial median per group was not significantly different.



(B) Distance to the spatial medians are not significantely different between beach environments. A visual representation of the average distances to the spatial median from **(A)**.

FIGURE 3.6 Beta diversity quantification using multivariate dispersion of beach sand and water bacterial populations.

To explore further structures which may underly the taxonomic makeup of these samples, permutational multivariate analysis of variance (MANOVA) was employed [101]. Permutational MANOVA is a non-parametric method for determining sources of variation within distance matrices. For taxonomic data, generating a distance matrix using an appropriate measure followed by permutational MANOVA can reveal what parameters are important for the distribution of samples within the multivariate space. Essentially, distances between samples are regarded as response variables and the covariates (parameters of interest) are the linear predictors. This technique is similar to analysis of similatiry (ANOSIM), but it has been shown to be more robust and less sensitive to dispersion effects [102]. Using the Bray-Curtis dissimilarity measure again, permutational MANOVA reveals a similar trend to what was observed for the richness and alpha diversity analysis (Table 3.5). The beach environment was the parameter that explained the most variation within sample distances, followed closely by the year. However, this analysis also uncovers the significance of the lake parameter and a type and year interaction, albeit these were small sources of variation relative to sample type or year. Removal of the Nickel Beach samples did not alter the significance of the permutational MANOVA test, although the percent of variance explained by environment increases. This is expected since the Nickel Beach samples show high similarity and would therefore result in an underestimation of the effect of this parameter. Overall, despite the spatial proximity of the paired sand and water samples, the variation imposed by the beach environment explains more of the bacterial taxonomic variance than samples from different lakes or even years.

TABLE 3.5 Permutational MANOVA of Bray-Curtis Dissimilarity

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Type	1	1.52	1.52	10.87	0.20	**0.0001**
Year	1	1.44	1.44	10.30	0.19	**0.0001**
Lake	1	0.44	0.44	3.15	0.06	**0.0120**
Type:Year	1	0.33	0.33	2.38	0.04	**0.0404**
Type:Lake	1	0.16	0.16	1.13	0.02	0.2966
Residuals	26	3.63	0.14		0.48	
Total	31	7.52			1.00	

So far, we have established through various means of diversity measures that significant differences underly the bacterial communities between beach sands and water. Although other effects may be in play, bacterial communities are more different between paired samples of water and sand than between lakes or years, suggesting that the comparative analysis of the microbial communities between beach environments is important. To extend this analysis, taxonomic and functional composition of the bacterial communities was explored with an emphasis on differences between these populations in the sand and water.

3.3.3 Taxonomic and Functional Composition

Taxonomic composition was initially assessed at the superkingdom level to identify the bacterial capture of the microbial communities. Although bacterial fractions were extracted during processing, all samples contained a small fraction of reads assigned to viral, eukaryotic, and archaic taxa as well (Figure 3.7). This is expected for complex microbial communities because the filtration process will not completely exclude non-bacterial taxa and free floating DNA from non-bacterial organisms. Furthermore, bioinformatic processing may introduce false non-bacterial detection due to inaccurate taxonomic classification, especially when utilizing protein sequences, which may exhibit strong conservation between divergent taxa. This is expected to occur at quite low levels however, and the large majority of samples were indeed comprised of a bacterial proportion of at least 90%. However, a subset of samples showed an enriched eukaryotic fraction, which was not necessarily reflected in the paired sand environment, suggesting a water-specific effect (e.g. 2.5.2013.LB.ER and 2.3.2013.LB.ER). This effect is also likely geographically and temporally isolated as the two samples with large eukaryotic fractions were the only samples from that same beach on the same day. The taxonomic constituents causing the eukaryotic enrichment were not a consequence of human contamination as diatoms and eukaryotic algal species made up the large majority of this enrichment. Since processing was conducted with bacterial characterization in mind, all further compositional analysis will focus solely on the bacterial constituents of these samples.

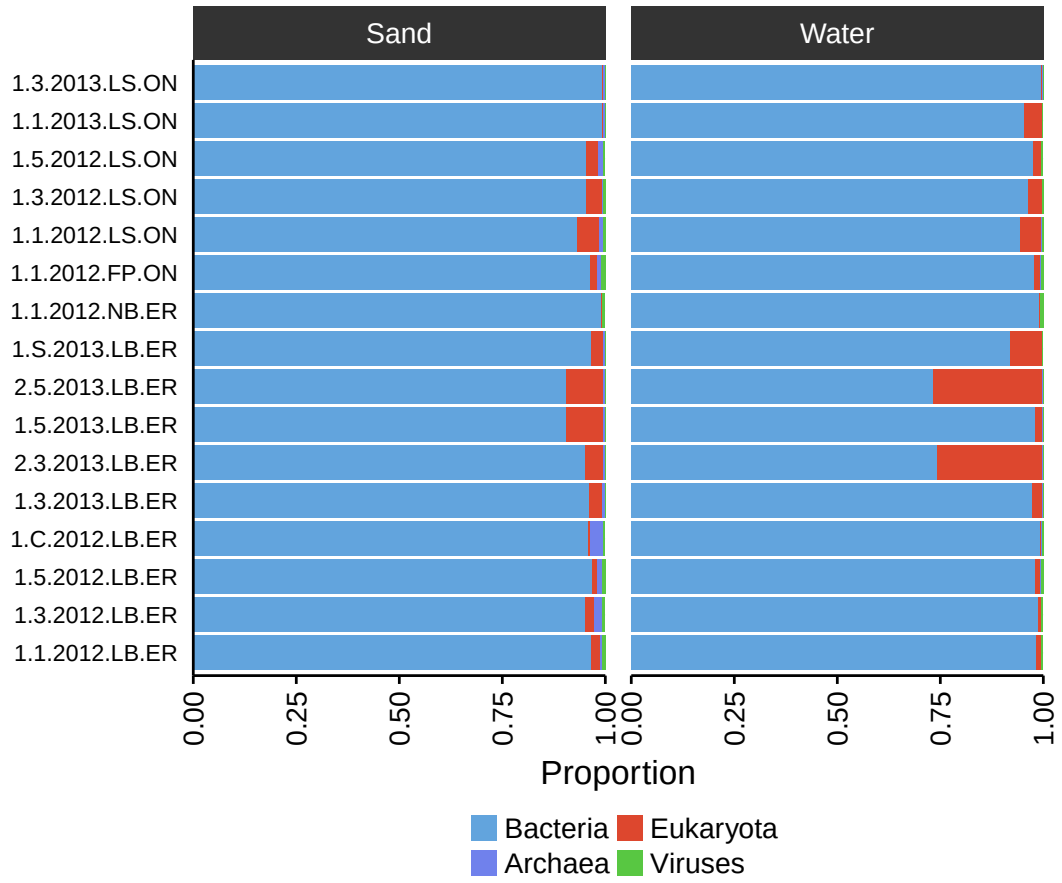


FIGURE 3.7 Bacterial capture of microbial populations was effective. Reads were aggregated to the superkingdom level to assess the proportion of reads assigned which belong to a bacterial origin.

Phylum compositions were examined next (Figures 3.8, 3.9). In both environments, a few phyla predominated the bacterial communities. In sand environments, Proteobacteria were the predominant phylum, although a large proportion of reads were also assigned to unclassified phyla. Other predominant phyla include the Actinobacteria, Bacteroidetes, Cyanobacteria, Firmicutes, Verrucomicrobia, and Planctomycetes. In comparison, the phylum distribution in the water show slightly different trends with a much more skewed and uneven distribution. Although Proteobacteria are the largest group again, the cumulative proportion of the Proteobacteria, Actinobacteria, and Bacteroidetes comprise up to 80 - 90% of all phyla in most samples, with a much smaller presence of the remaining phyla, indicating a decreased diversity of taxa at the phylum level. The diversity at the phylum rank is indeed significantly greater in the sand than in the water (paired t-test, $p = 3.93 \times 10^{-4}$), which supports the observation that the bacterial communities of these freshwater environments are dominated by fewer phyla

than those in the sand and also agrees with the species level diversity. The diversity at the phylum level further extends to unique phyla present in the sand but not the water (Table 3.6). Phyla were considered unique to an environment if presence was detected in at least 9 of the 16 samples from that environment but not detected in any samples of the corresponding environment. Interestingly, all unique phyla were present only in the sand, further providing evidence for an enriched diversity and richness of bacterial communities in the sand than in the water. Further, many of these unique phyla correspond to uncultured organisms or poorly characterized phyla grouped by organisms detected solely through metagenomic methods, suggesting a plethora of unexplored bacterial constituents present in the sand that are absent in the water. This corresponds with the high detection of taxa which correspond to unclassified phyla in the sand as well (Figure 3.8). The Aquificae were unique to the sand environment, and typically exhibit persistence in thermophilic environmental conditions and optimal growth temperatues at $65°$C or higher [103]. Although they are aerobic, they can employ anaerobic respiration through the use of thiosulfate or sulfur and are often found in sulfur pools [104]. The Aquificae detected in the sand all belonged to the Aquificaceae, Desulfurobacteriaceae, and Hydrogenothermaceae families, and the taxonomic resolution was not specific enough to determine the species. The candidate division NC10 are a known aquatic group of organisms initially detected through 16s rDNA sequencing, and includes an anaerobic methane oxidizer [105]. The Latescibacteria are another recently identified group with presence in anoxic conditions and sediment environments [106]. The Latescibacteria additionally possess many functions relating to the processing of organic polymers which arise from algal sources. Overall, many of these unique phyla have not been detected in sand habitats previously, and possess niche functions which may lend to the unique bacterial populations present in this environment.

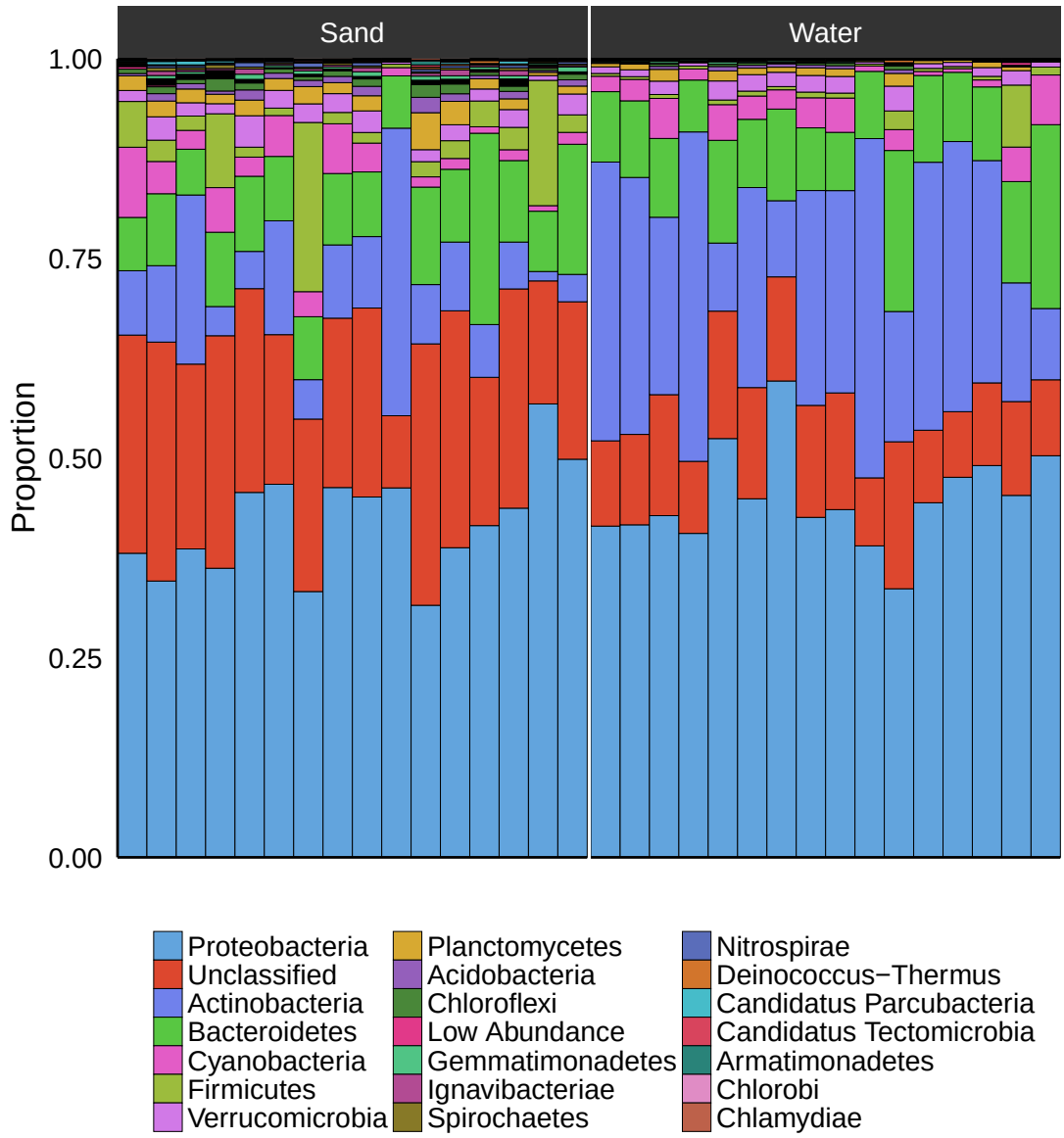


FIGURE 3.8 Phylum compositions of beach sand and water sites. Reads assigned to bacterial taxa were aggregated to the phylum level and plotted as proportions. Reads assigned to a bacterial taxa whose phylum classification was unknown are aggregated and labelled as unclassified. Sites are distinguished by beach environment.

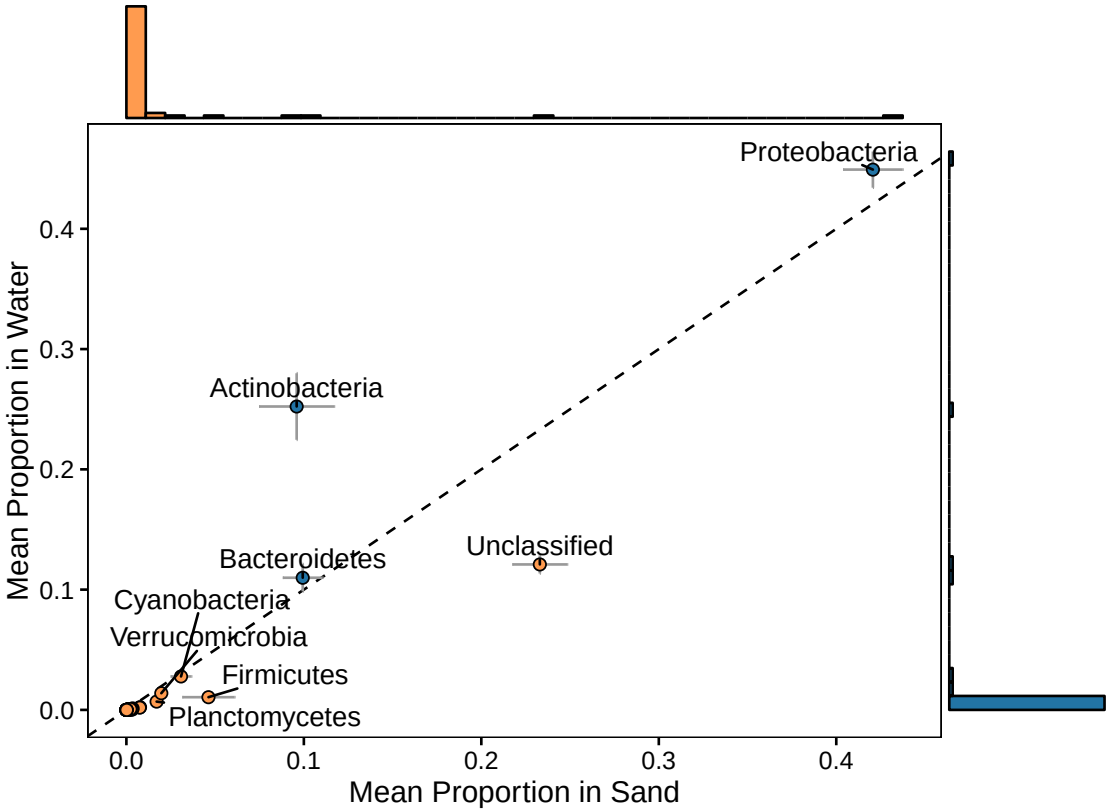


FIGURE 3.9 Mean phylum proportions of bacterial populations in beach sands and water. The mean proportion of phyla in both sand and water are plotted against each other with standard error of the mean proportion indicated for each environment. The marginal histogram describes the frequency of phyla at the given proportions.

TABLE 3.6 Phyla Unique to Habitat Type

	Number of Samples		Total Reads	
Phylum	Sand	Water	Sand	Water
Aquificae	14	0	1259	0
candidate division NC10	14	0	1244	0
Latescibacteria	13	0	1829	0
candidate division Zixibacteria	11	0	902	0
Candidatus Microgenomates	9	0	2868	0
Candidatus Aminicenantes	9	0	1343	0
Nitrospinae	9	0	737	0
Candidatus Hydrogenedentes	9	0	679	0
Deferribacteres	9	0	585	0

* Unique phyla determined by presence in at least 8 samples within one type and absence in alternate type

Interestingly, broad level variations in phylum composition did not translate to similar variations in functional capacity as determined by the SEED subsystem classifications (Figures 3.11, 3.12). Between samples and environment types, the broad level functional capacity does not exhibit much variation. Functions pertaining to macromolecule processing and metabolism are detected with the greatest abundance across all samples, with more specialized functions relating to element acquisition or metabolism occurring in smaller abundances. This suggests that despite marked variations in phylum composition within and between environments, functional capacity is maintained. However, it is possible that the hierarchical level used for SEED classification was not specific enough to visualize variations between samples or environments. Since classifications were aggregated to level two of the SEED subsystem, variation between samples or sample types may be minimized due to a reduced resolution, causing an artificial homogeneity of functional capacity between samples. To ensure that the lack of variation across samples was not a potential artifact of the methodology, a non-related sample was put through the same pipeline of analysis as a control and functional capacity was examined (see C (control) column in Figure 3.11). If this sample also shows a similar functional profile, this would indicate that the level at which functional capacity is being examined is not specific enough to detect variation. However, if the functional profile is different, this would suggest that the lack of variation seen in the beach environments is representative of the functional profiles. Although similar functional categories were detected in the control dataset, the relative abundances of these functions suggest a divergent functional profile compared to the beach samples. As well, this is further confirmed via PCoA, which highlights the distance between the beach samples and control sample as well as the lack of sample clustering based on environment (Figure 3.12b). This finding suggests that the level of the SEED subsystem is still specific enough to detect variation between samples with different functional profiles, and that there exists a lack of variation in functional capacity between the sand and water environments at this level.

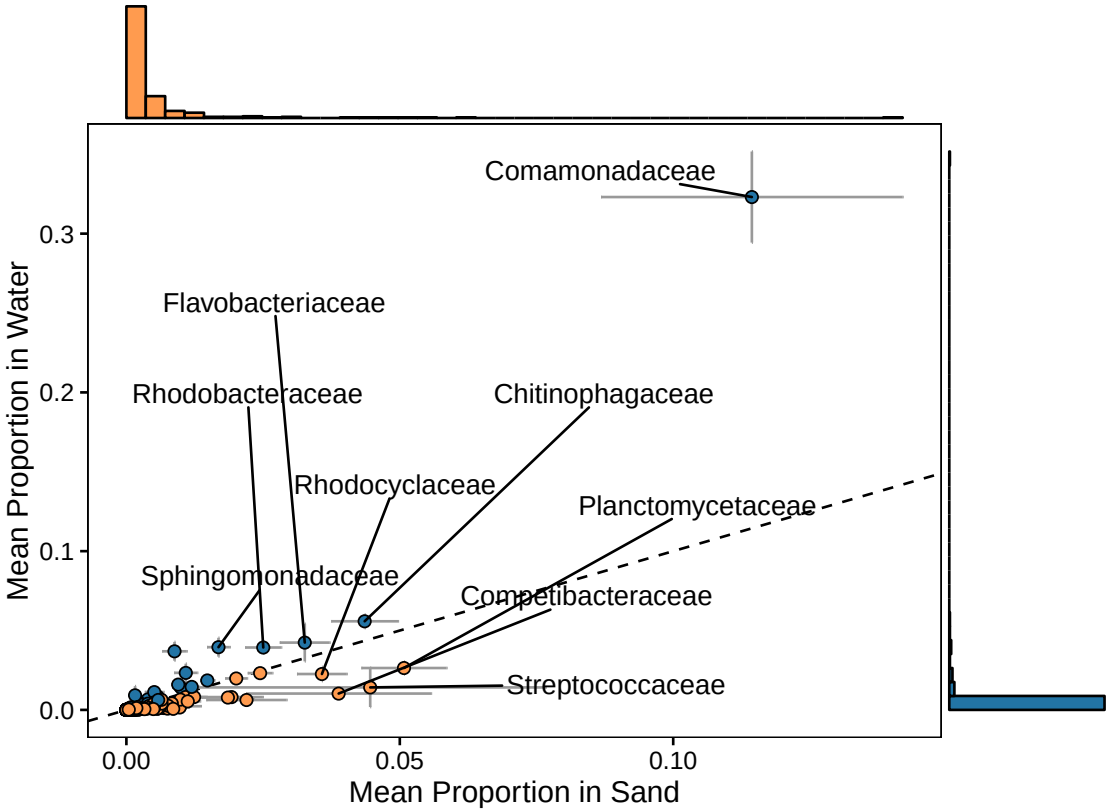


FIGURE 3.10 Mean proportions of bacterial taxa at the family level in beach sands and water. The mean proportion of families in both sand and water are plotted against each other with standard error of the mean proportion indicated for each environment. The marginal histogram describes the frequency of phyla at the given proportions. Families labelled as unclassified were removed prior to determining mean proportions and are not plotted.

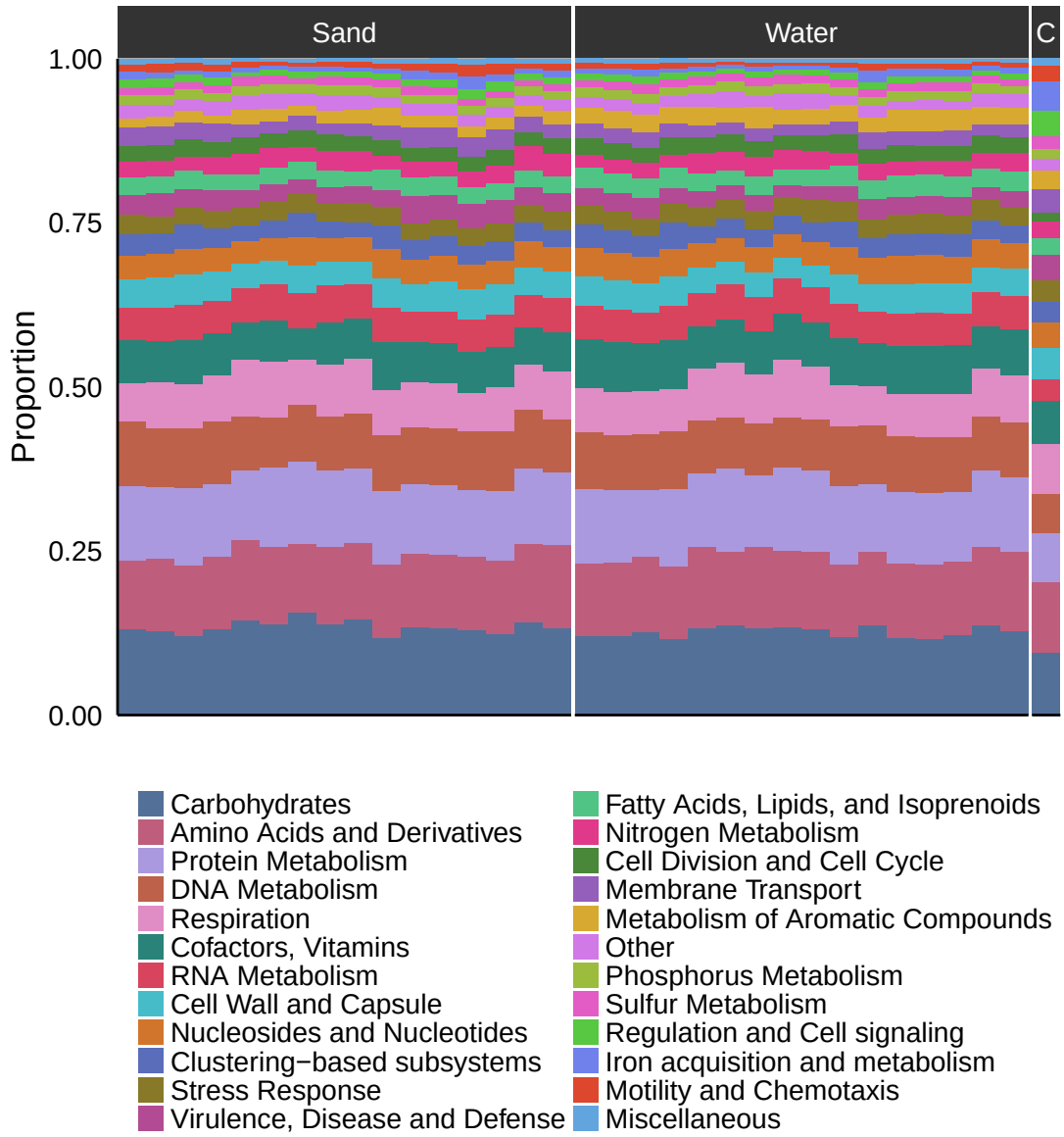
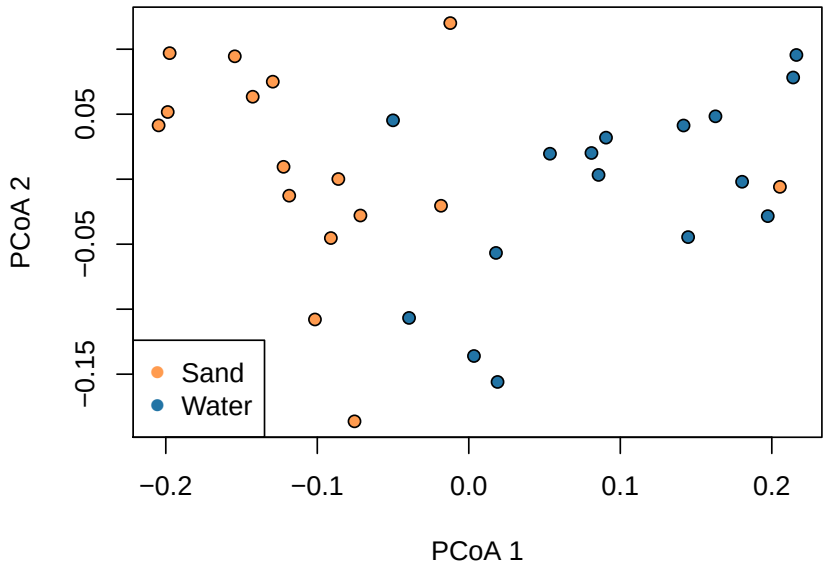
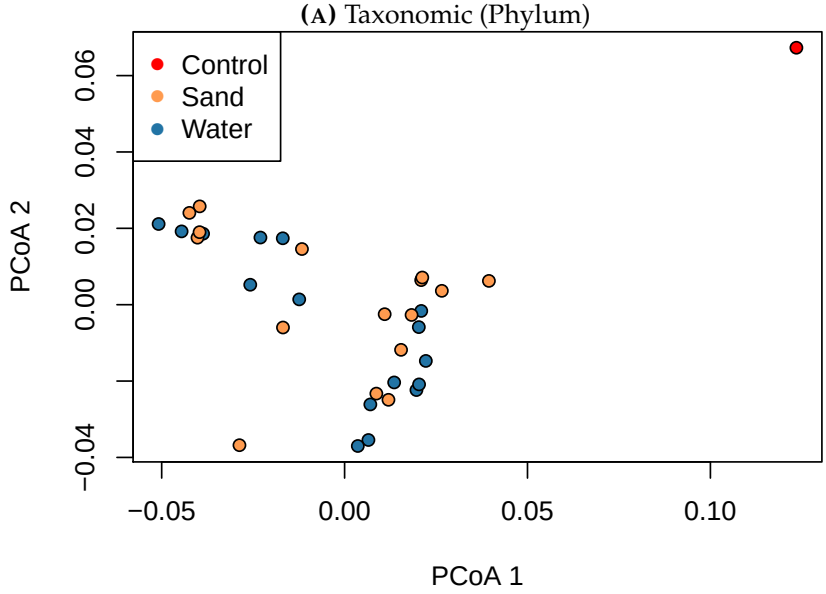


FIGURE 3.11 Functional capacity is stable between samples and environments. Functional classification to the SEED subsystem hierarchy aggregaetd to level two is plotted as proportions of the number of reads with a funcitonal assignment for each sample. Samples corresponding to sand and water sites are separted, with the 'C' site corresponding to a control shotgun metagnome of sputum sample from Cystic Fibrosis patient (accession: SRX1655756).



(A) Taxonomic (Phylum)



(B) Functional (Level Two)

FIGURE 3.12 Principal coordinates analysis reveals variable taxonomic composition but stable functional capacity. PCoA using Bray-Curtis dissimilarity measures of **(A)** taxonomic composition at the phylum level and **(B)** functional capacity at level two of the SEED subsystem hierarchy.

A breakdown of the Proteobacteria was examined next because of their dominance across all samples (Figure 3.13). Betaproteobacteria, Alphaproteobacteria, and Gammaproteobacteria were the most dominant groups of Proteobacteria across both environments. At the family level, Comamonadaceae belonging to the Betaproteobacteria occurred at the highest abundance and were seemingly more abundant in the water (Figure 3.10). The Comamonadaceae are aerobic motile organisms and are therefore expected to be more abundant in the water, although their relatively high abundance in the sand suggests that bacterial transfer between the water and sand is occurring. Other high abundance families include Flavobacteriaceae, Rhodobacteraceae, Streptococcaceae, and Planctomycetaceae.

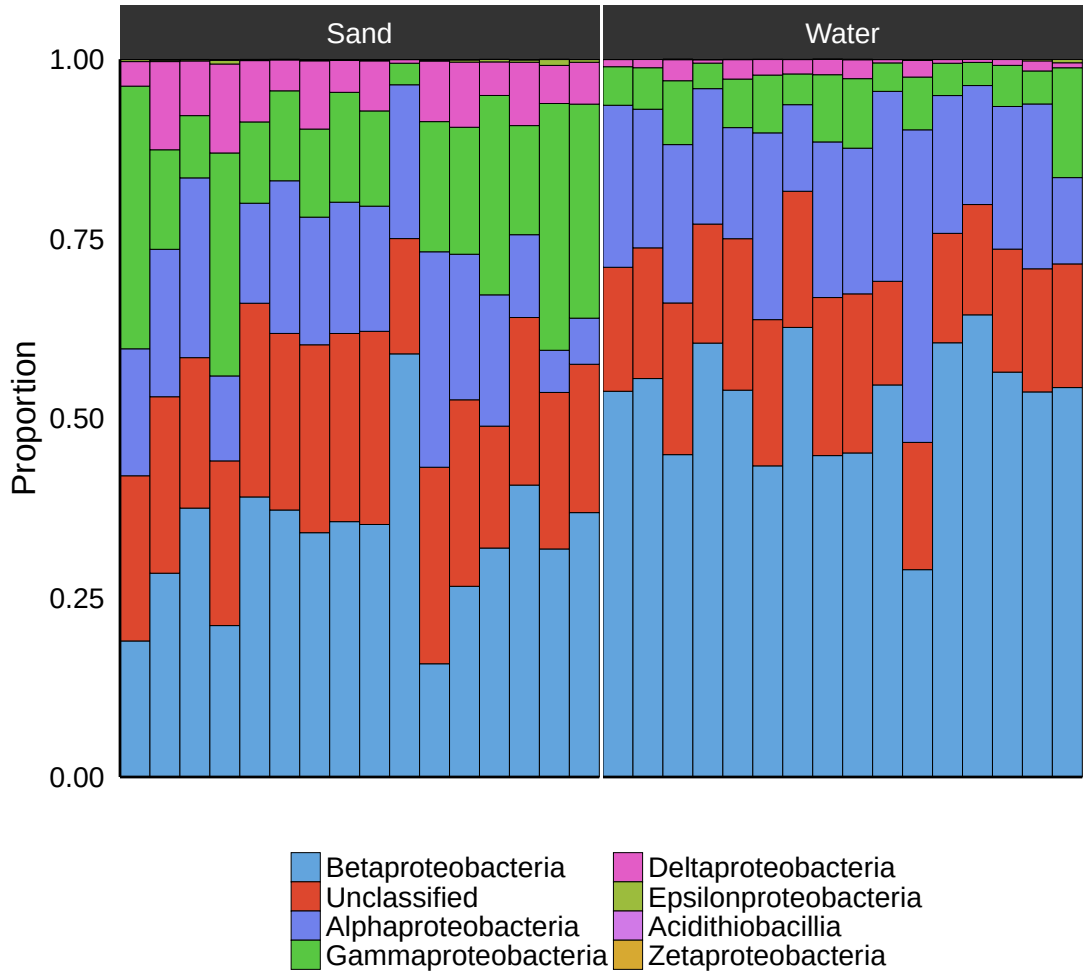


FIGURE 3.13 Alpha diversity measured by the Shannon index between the two sample types, sand and water. The diversity measure was captured at the species level and excludes taxa assigned to higher taxonomic ranks. Standard error of the mean are represented by the error bars

3.3.4 Differential Abundance of Taxonomic and Functional Features

Broad level compositional analysis is important for identifying general trends, but higher resolution analysis is required to determine significantly important features between environments and for the application of this analysis to pathogen detection. Differential abundance, a term analogous to differential expression in RNA-seq experiments, was used to identify significant differences in taxonomic and functional features between environments. An important first step in the detection of differentially abundant features is count normalization, since sequencing depth is not equivalent across all samples. Using proportions has been shown to lead to erroneous conclusions about differential abundance (and expression), since it does not account for the heteroscedasticity that is often present in microbiome and metagenomic data [100]. Therefore, models which account for differences in library size and biological variance have been developed in an attempt to more accurately describe count data. Comparative analyses of various models (as implemented in R packages) for microbiome data were conducted by McMurdie and Holmes [100] and Weiss et al. [107] to investigate the sensitivity and power of different statistical models. Although packages like DESeq2 and edgeR were developed initially for RNA-seq experiments, both studies found that these packages performed exceptionally well for the detection of differentially abundant features in microbiome data, with DESeq2 having the strongest power [107]. metagenomeSeq is a package specifically developed for microbiome data, however, it exhibited poorer performance than other packages, and showed high rates of false positives, especially with a small number of samples (< 50 samples per group) [107]. An additional method for differential abundance is the use of non-parametric testing of differences in means (Mann-Whitney U test), however, Weiss et al. [107] suggest the use of non-parametric methods when the number of samples exceeds 50. Based on these studies, DESeq2 was employed for testing differential abundance.

Significant differences in many phyla are detected between beach sands and waters (Figure 3.14). Eighteen phyla were enriched in the sand, but apart from Firmicutes, were present in very low abundances in both environments. The marked enrichment of Actinobacteria, Bacteroidetes, and Proteobacteria in the water is in contrast to the enrichment of many low abundance phyla in the sand, adding support to the notion of a more diverse bacterial population in the beach sands. The differences in phylum composition extend down to significant differences at the species level, which seem to align in proportion to the abundance of the phyla (most differentially abundant species are Proteobacterial, followed by Actinobacteria etc.; Figure 3.15). The increase in Actinobacteria seen in the water is peculiar because Actinobacteria were previously thought to be typical soil-dwelling bacteria. However, the increase in Actinobacteria in the water corresponds with the enrichment of fourteen Actinobacteria species which belong to a group of planktonic organisms prevalent in freshwater environments (Table 3.7). These organisms seem to be related to a highly abundant group of freshwater Actinobacteria which show presence in many freshwater environments [108, 109].

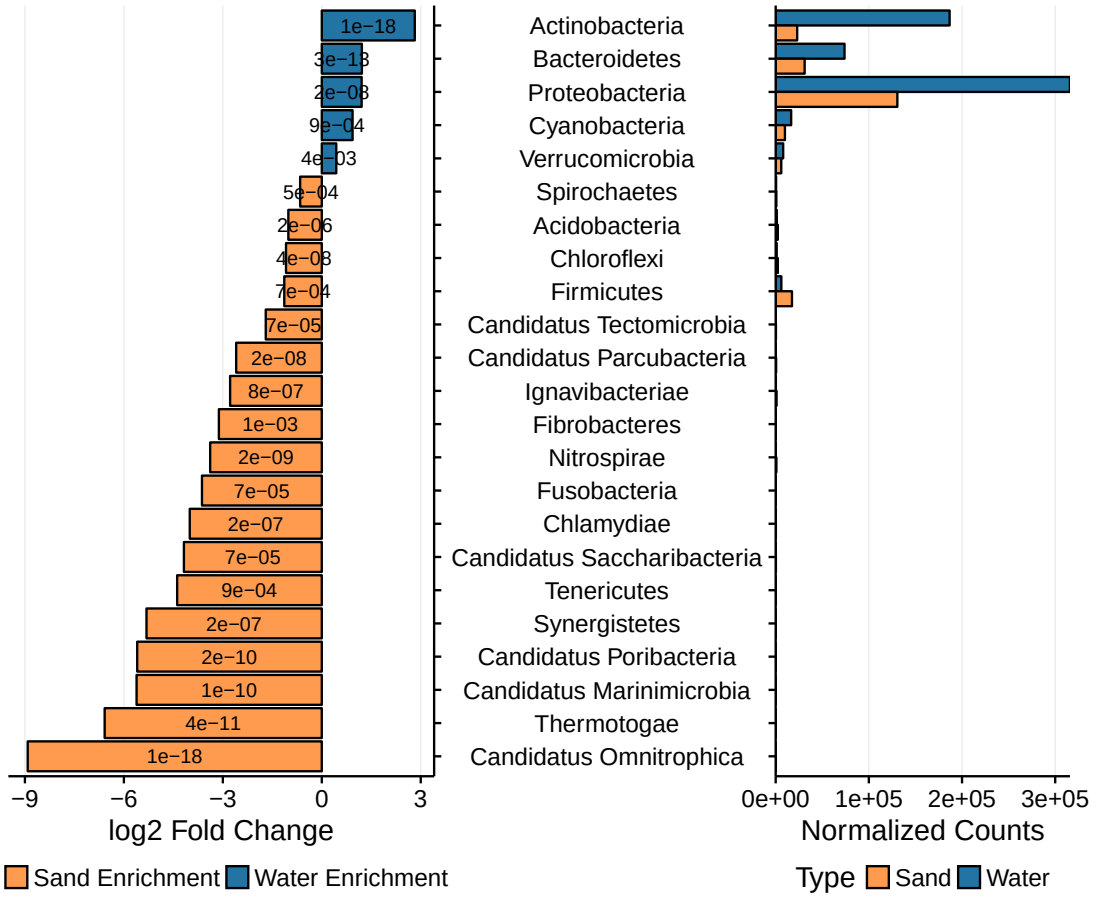


FIGURE 3.14 Many phyla are differentially abundant between beach environments. Differential abundance of phyla was determined by DESeq2 using a paired model. log2 fold changes are indicated on the left with the associated adjusted p-values of significance (cutoff of 0.01). The normalized counts for each phyla in each environment are plotted on the right.

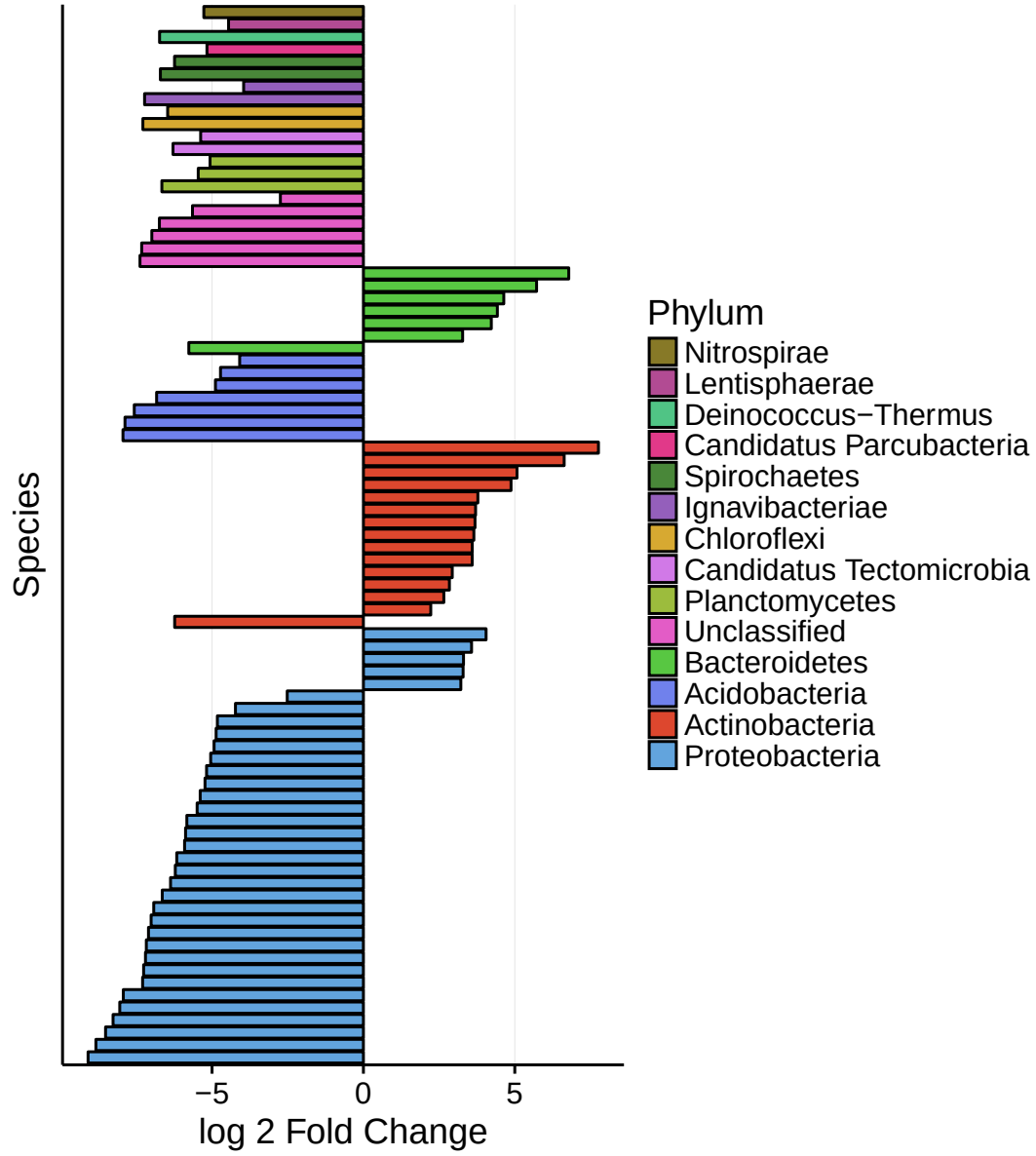


FIGURE 3.15 Differentially abundant species between beach environments. Differential abundance of species was determined by DESeq2 using a paired model. log2 fold changes (l2FC) are indicated, with negative l2FC signifying enrichment in the sand and positive l2FC signifying enrichment in the water. Species are coloured according to their phylum classification and significance was determined at an adjusted p-value of 0.05.

TABLE 3.7 Actinobacteria Species Enriched Between Beach Environments

Species		Normalized Mean	log2FoldChange	lfcSE	padj
actinobacterium	SCGC AAA023-D18	3480.09	2.66	0.78	0.0045
	SCGC AAA027-J17	1644.75	2.93	0.91	0.0076
	acAcidi	1271.97	2.22	0.70	0.0087
	SCGC AAA028-I14	1107.09	3.78	1.00	0.0013
	SCGC AAA278-O22	940.37	3.69	0.97	0.0013
	SCGC AAA027-M14	782.13	3.59	1.05	0.0042
	SCGC AAA044-N04	508.78	3.71	1.13	0.0069
	SCGC AAA024-D14	494.75	3.59	1.13	0.0090
	SCGC AAA044-D11	377.80	3.65	1.10	0.0063
	SCGC AAA027-D23	270.05	2.84	0.90	0.0094
	SCGC AAA023-J06	155.01	4.88	1.26	0.0010
	SCGC AAA028-A23	42.79	6.63	1.53	0.0002
Candidatus Aquiluna	sp. IMCC13023	72.87	5.07	1.61	0.0095
uncultured actinobacterium		36.98	7.76	1.79	0.0002
Conexibacter woesei		33.85	-6.23	1.42	0.0002

A large number of the Proteobacteria species enriched in the sand belonged to the sulfur- and sulfate-reducing Deltaproteobacteria (38 out of 56), and belong to genera with functions associated with reduction of sulfur-containing compounds (Table 3.8). The remaining species that were identified as enriched in the sand mostly belonged to uncharacterized bacterium with non-informative identifiers (e.g. bacterium UASB14 and UASB270, identified previously in anaerobic sludge blankets through 16s rDNA sequencing [110]) or taxa labelled as uncultured bacterium.

TABLE 3.8 Frequency of Proteobacteria Orders of Species Enriched in Sand

Order	Frequency
Bdellovibrionales	2
Burkholderiales	3
Chromatiales	1
Desulfarculales	1
Desulfobacterales	9
Desulfovibrionales	1
Desulfuromonadales	6
Gallionellales	1
Methylococcales	3
Myxococcales	6
Rhodocyclales	4
Sulfuricellales	1
Syntrophobacterales	7
Thiotrichales	1
Unclassified	10

Although this analysis cannot make conclusions about the immediate response of the bacterial communities to the environment by examining the expression of genes, assessment of the functional capacities and functional enrichment of relevant functions is possible by examining abundance of genes with relation to categorized functions. This information can reveal what the communities have adapted to and what functions are more prevalent due to changes in taxonomic constituents. Differential abundance in many functions relating to metabolic process were detected (Figure 3.16). This supports recent findings by [111] that species compositions vary in accordance to biochemical metrics such as pH, salinity, and nutrient load/density. Importantly, functions pertaining to capabilities expected in tough environmental conditions like those experienced in the sand such as sporulation and spore coat genes, acid stress, and sigmaB stress regulation are enriched in this environment. The enrichment of flavohaemoglobin (nitric oxide dioxygenase) and nitrosative stress genes in addition to nitrite reductase and nitrate/nitrite ammonification genes also suggests the abundance of nitrogen-containing compounds and their potential cause of nutrient stress in the bacterial community. The anoxic conditions of sands also seem to enrich growth of anaerobic bacteria and persistence of genes associated with anaerobic functions such as anaerobic degradation of aromatic compounds and sulfate reduction-associate complexes. The enrichment of sulfate reduction-associated complexes in the sand also coincides with the enrichment of the Deltaproteobacteria, and sulfur- and sulfate-reducing organisms enriched in the sand. Enrichment of cytochrome c oxidase biogenesis in the water communities also further lends support to the greater capacity of aerobic respiration in the water, whereas anaerobic respiration likely dominates in the sand.

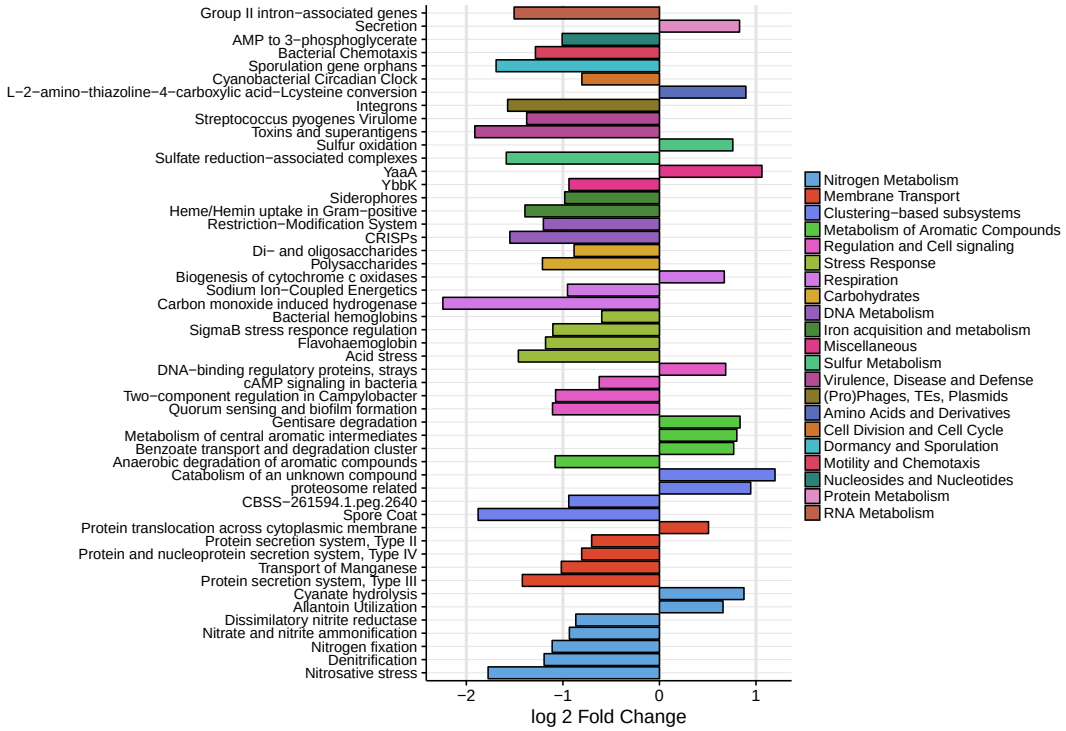


FIGURE 3.16 Functions with differential capacity between beach environments. Differential abundance of functions was determined by DESeq2 using a paired model. log2 fold changes (l2FC) are indicated, with negative l2FC signifying enrichment in the sand and positive l2FC signifying enrichment in the water. Functions depicted are in level three of the SEED subsystem hierarchy and are coloured by their broader level two categorization and significance was determined at an adjusted p-value of 0.01.

3.3.5 Pathogens and Indicator Bacteria

An important goal of this work is to establish the shotgun metagenomic approach as a viable solution for the detection of pathogens and FIB within freshwater beach environments. Pathogens and FIB were selected based on their relevance and potential impact to public health in recreational waters and previous detection in beach sands and waters (Table 3.9). Many of the pathogens were selected from the work of Whitman et al. [112], who discuss the prevalence of microbial pathogens in the sand of beaches. Since some studies use molecular markers which could not identify at the species level, all species belonging to that genera that were detected were examined. There was large variability in the detection of various pathogens and FIB (Table 3.10). While some pathogens and FIB such as *Aeromonas* spp., *E. coli*, *Salmonella enterica*, and *Pseudomonas aeruginosa* were detected in almost all samples, others were not detected at all (e.g. many *Campylobacter*). Additionally, although many reads were assigned to pathogenic species in the sand and

water of the same site, there were many instances of asymmetric detection solely in one environment of a beach and not the other (e.g. *Vibrio* spp. and *Campylobacter lari*). Many of the *Vibrio* species examined were detected in both the beach water and sand or just the beach water, but rarely detected solely in the sand at any given site. This was also the case for *Staphylococcus aureus*.

TABLE 3.9 Pathogens and Indicator Bacteria Of Interest

Taxa	Pathogenicity	Previous Detection in Beach Environments	Reference
Aeromonas spp.	Pathogen (GI)	Marine beach water Foreshore sand	[89, 113]
Bacteroides spp.	FIB / Opportunistic	Marine/Freshwater beach water Marine beach sand Freshwater beach water	[82, 114–116]
Campylobacter spp.	Pathogen (GI)	Marine/Freshwater beach sand	[90, 91, 117]
Clostridium perfringens	Pathogen (GI; enterotoxins)	Marine beach water Marine beach sand	[113, 118, 119]
Escherichia coli	FIB / Pathogenic strains	Marine/Freshwater beach water Marine/Freshwater beach sand	
Legionella pneumophila	Pathogen	Marine waters	[120]
Pseudomonas aeruginosa	Pathogen	Marine beach water Marine beach sand	[118]
Salmonella spp.	Pathogen (GI (non-typhoidal)	Marine beach sand	[91]
Shigella spp.	Pathogen (GI)	Marine beach water Marine beach sand	[121]
Staphylococcus aureus	Pathogen	Marine/Freshwater beach water Marine beach sand	[118, 122]
Vibrio spp.	Pathogen (GI)	Marine beach water Marine beach sand	[113, 123]

TABLE 3.10 Pathogen Detection

Taxa	Detection In		
	Both	Sand	Water
Aeromonas hydrophila	15	1	0
Aeromonas salmonicida	15	1	0
Aeromonas veronii	16	0	0
Bacteroides fragilis	8	2	2
Bacteroides helcogenes	14	0	1
Bacteroides salanitronis	3	3	3
Bacteroides thetaiotaomicron	1	4	0
Bacteroides vulgatus	5	2	5
Campylobacter coli	2	0	3
Campylobacter concisus	0	0	0
Campylobacter curvus	0	0	0
Campylobacter fetus	0	0	1
Campylobacter hominis	0	1	1
Campylobacter jejuni	0	3	4
Campylobacter lari	0	1	6
Clostridium perfringens	9	0	2
Escherichia coli	16	0	0
Legionella pneumophila	10	1	1
Pseudomonas aeruginosa	16	0	0
Salmonella bongori	0	1	3
Salmonella enterica	16	0	0
Shigella dysenteriae	0	1	1
Shigella sonnei	0	0	0
Staphylococcus aureus	3	1	4
Vibrio alginolyticus	0	0	0
Vibrio anguillarum	1	0	4
Vibrio antiquarius	0	1	3
Vibrio campbellii	9	2	4
Vibrio cholerae	5	1	5
Vibrio furnissii	4	0	8
Vibrio nigripulchritudo	1	2	6
Vibrio parahaemolyticus	3	1	7
Vibrio sp. EJY3	0	1	5
Vibrio tasmaniensis	2	0	6
Vibrio vulnificus	6	1	4

In addition to variation in detection of the numerous pathogens investigated, the number of reads assigned to the various pathogens and FIB differed greatly (Figure 3.17). In the sand, *Aeromonas veronii* and *E. coli* were detected in the largest abundance. Similarly, a large number of reads were assigned to *E. coli* in the water, however the

assignment of reads to *A. veronii* was less than in the sand, but not significantly so ($p > 0.05$). *Vibrio* spp. were detected in low abundances and most showed significant enrichment in the water ($p < 0.05$). The sand had significantly higher counts of *A. hyrophila* ($p = 2.87 \times 10^{-2}$) and *P. aeruginosa* ($p = 9.46 \times 10^{-4}$). Although *S. enterica* was detected in all samples, the number of reads assigned was relatively low and was not significantly different between the two beach environments ($p > 0.05$).

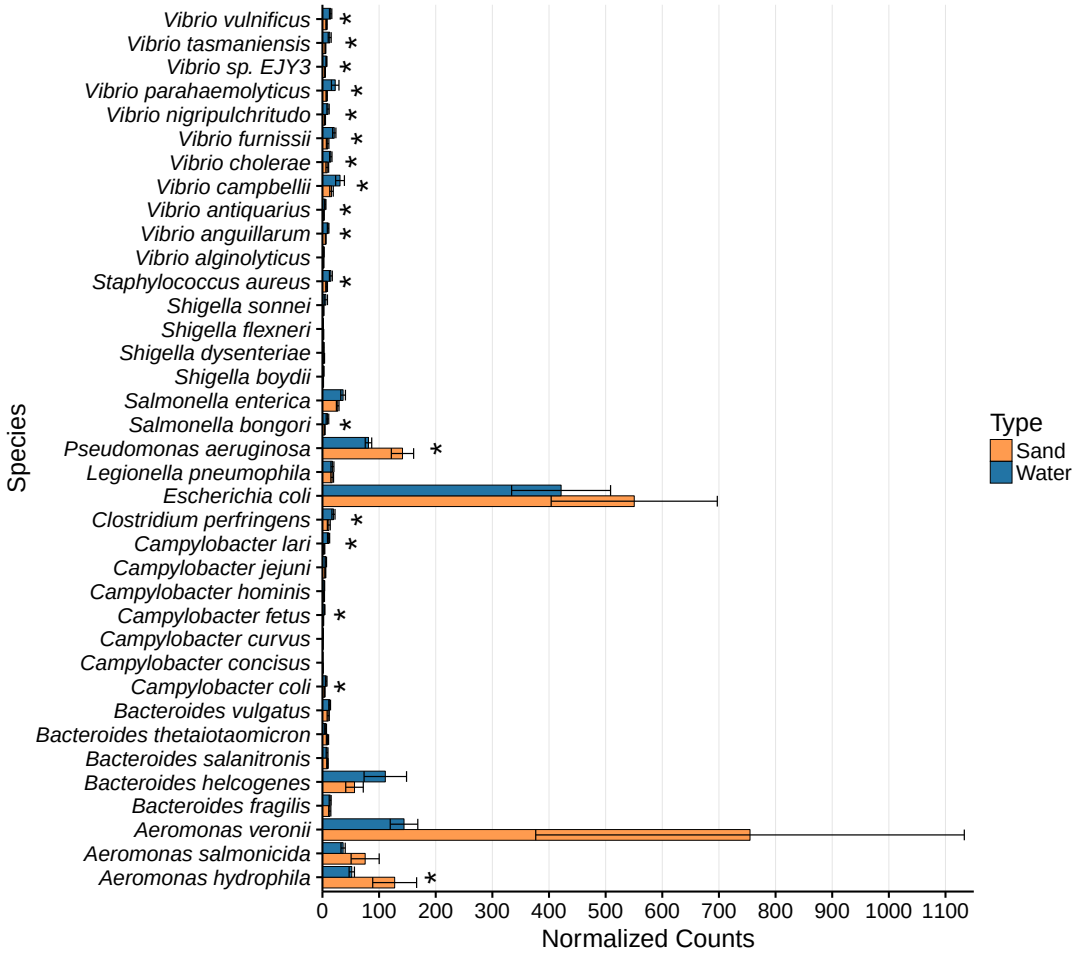


FIGURE 3.17 Reads assigned to pathogens are present in both beach environments. Reads were classified using CLARK as described in the methods. Counts were normalized using DESeq2's size factor estimations and mean normalized counts for each environment were plotted with the error bars indicated standard error of the mean. Significance between environments was determined using DESeq2's negative binomial model testing.

3.4 Discussion

This work extends the understanding of bacterial populations in freshwater beaches and also provides insight into the similarities and differences of these populations between beach environments. To our knowledge, there are no other investigations of freshwater beach environments which utilize a shotgun metagenomic approach, and this work can serve as a preliminary investigation of these communities using this method. Our findings suggest that the beach environment is an important parameter which strongly influences ecological measures and taxonomic composition of bacterial populations. Bacterial communities are more distinct between beach sands and waters than between years or lakes, and this effect was not dependent on different lakes or years. However, since only two lakes and two years were examined, this analysis would benefit from an expansion to more lakes and years to accurately assess the relative importance of the beach environment compared to large scale spatial or temporal effects.

Previous work by Cui et al. [124] examined the spatial effects imposed by different beach environments on the bacterial communities of marine beaches in Hawaii using 16s rDNA sequencing. Their work encompassed the analysis of beach water and three beach sand environments; nearshore sand, foreshore sand, and backshore sand. The backshore sand in their work is analogous to the location of the sand pores generated in our work. Their investigation revealed similar trends to our findings. Mainly, rarefaction curves did not reveal differential saturation of bacterial constituents between backshore sand and beach waters, and richness and diversity were significantly elevated in backshore sands compared to beach waters. Additionally, ordination revealed strong clustering of samples originating from backshore sands distinctly from beach waters. In general, the findings described in this work seem to coincide well with previous analysis of beach sand and water bacterial populations, and also suggests that these trends are valid in both marine and freshwater beaches. Their findings also suggest that bacterial differences between beach waters and backshore sand do not apply to sand which is wave-washed, and this also likely is true in freshwater environments, although this would need to be verified by examining pore samples from varying sand environments. Our results slightly differ, however, due to the detection of many more taxonomic features using our methods. This could suggest that a shotgun metagenomic approach is more sensitive in capturing the microbial populations, but could also be due to a genuine depression in taxonomic richness in marine environments than freshwater environments. However, 16s analysis has been known to have crude taxonomic resolution [37] and this might be the reason for a lower richness detected in the work by Cui et al.

The use of sand pore samples in this work as a proxy for sand habitats was employed due to its standard use in municipal sampling in the Niagara Region. Cui et al. used deionized water to release bacterial cells, while the use of sand pores works under the assumption that water flowing through the sediment will release bacterial cells.

Both similarly work on the same principle, and although the sand pore method has been employed before and established in literature (e.g. [112]), there may be a potential bias imposed when the capture of one environment (the sand) is dependent on the other (the water). This could potentially depress the distinction between the two environments due to the dependency of the sand pore samples on the beach water and may have impacted the comparative investigation of these two environments. For future studies, it would be important to assess the validity of this method within the context of comparing beach sand and water habitats.

Significant differences were seen between the beach sand and water in terms of taxonomic composition at a broad level. These differences included the decreased diversity in the water as Proteobacteria and Actinobacteria exhibited greater abundance, while decreased abundance of the predominant phyla in the sand permitted increased abundance of less dominant phyla. Although Proteobacteria, Actinobacteria, and Bacteroidetes were enriched in the water, they were the dominant phyla along with the Firmicutes detected in the sand, a finding which supports 16s rDNA analysis of beach sands discussed by Whitman et al. [112]. However, Whitman et al. mention that the most abundant families detected in the sand were Rhodobacteraceae, Flavobacteriaceae, Flammeovirgaceae, and Campylobacteraceae. In our work, high abundance was also seen for Rhodobacteraceae and Flavobacteriaceae, but Comamonadaceae were the most dominant family, in contrast to their findings. Their analysis consisted of two marine and one freshwater site however, and the inclusion of marine sand environments may result in differential family abundances.

Enrichment in the sand of many Deltaproteobacteria which contain known sulfur- and sulfate-reducing organisms, in addition to the enrichment of genes associated with these functions and other anaerobic processing functions suggests the presence of anoxic conditions in the sand. As well, enrichment of functions relating to sporulation, nitrosative stress, acid stress, and flavohaemoglobin indicates harsh environmental conditions requiring specific functions for suitability in this environment. Analysis into the geochemical properties of the beach sands examined may reveal interesting patterns regarding various elemental and nutrient loads which may explain the enrichment of these organisms and functions.

Future analysis would benefit from deeper sequencing efforts with the intention of examining functional capacity more precisely. The lack of functional saturation suggests that the functional profiles are not complete, and introduces difficulty when attempting to assess the comparative functional capacity of these environments. Deeper sequencing could reveal unique metabolic features as many metagenomic sequencing datasets have done before (e.g. see [125, 126]). Deeper sequencing efforts may also be beneficial for extracting information regarding what taxa provide which functions. Although taxonomic and functional assignments can be traced back to individual reads in our dataset, the overlap of reads assigned to both a taxa and function were minimal, making it difficult to assess whether functional capacities coincide directly with taxonomic diversity. With deeper sequencing, changes in sulfur-reduction associated

genes may be directly correlated with changes in Deltaproteobacteria abundance, for instance.

3.4.1 Relavence to Public Health

This work has served as a means of introducing the potential use of shotgun metagenomics for implementation in water quality monitoring programs. The high confidence in detection of reads originating from various pathogens suggests that this method may be important in heading towards a more comprehensive means of water quality monitoring. Using CLARK and the unique k-mer principle for taxonomic classification, reads were assigned with high confidence. Manual nucleotide blast of random reads assigned to *E. coli* revealed 100% sequence identity to the *E. coli* and no other species (results not shown). A drawback however, is that many pathogens are specific strains of species, and these metagenomic methods have not advanced enough for the detection of pathogenic strains. Manual curation of the nucleotide blast results revealed that many of the genomic regions were non-discriminatory between various strains of species. With deeper sequencing, capturing genomic regions which are differ between specific strains may be possible, and may overcome this drawback. Taxonomic classification methods must also be developed for this analysis to be feasible, but in the interest of performance gains, even the newest classification programs may consider strains of species as redundant and omit the variability seen between strains [127]. Therefore, with increased sequencing depth and classification programs specifically developed for this purpose, strain identification may be possible, although the feasibility may be questioned.

Additionally, for adequate assessment of pathogen presence in these environments, further analysis is required to establish a correlation between the detection of reads in a metagenomic dataset and the true abundance. Standardization of sampling, processing, and sequencing must be developed and correlated with traditional methods before metagenomic methods may be implemented to monitor the pathogen loads of these beaches. This idea has been discussed previously by Jones et al. [128] within the context of human microbiome analysis and whole genome shotgun metagenomic sequencing, but should be extended to metagenomic analysis of environmental samples as well. Their findings also suggest that spike-in controls provide an adequate means of organism quantification in clinical samples. Spike-ins of *Shewanella oneidensis* in known concentrations to stool samples and quantification via qPCR correlated strongly with the relative abundance detected via metagenomic methods. This idea can be extended to water quality monitoring programs, and in conjunction with the comprehensive capture of the microbial populations using a shotgun metagenomic approach, may be a useful advance in water quality monitoring methods.

A call for more comprehensive and unified methods of pathogen detection in water bodies is not a new idea, but seems to have stagnated in the literature [83]. The employment of a metagenomic approach seems to be a viable answer to the issues surrounding

culture-based methods of water quality monitoring. With advances in sequencing technologies to reduce time and costs, increased sensitivity in the detection of pathogens, and adequate quantification standards, this approach could be utilized in water quality monitoring programs.

Appendix A

Chapter 2 Supplement

A.1 Sample Metagenomic Communities

TABLE A.1 Low Complexity Community

Rel. Ab.	Species	Taxon ID
9	Polaromonas naphthalenivorans CJ2	365044
1.0	Sulfuricella denitrificans skB26	1163617
1.0	Acidithiobacillus ferrivorans SS3	743299
0.8	Xanthobacter autotrophicus Py2	78245
1.0	Methylotenera versatilis 301	666681
0.8	Clostridium beijerinckii NCIMB 8052	290402
1.0	Dechloromonas aromatica RCB	159087
⋮	⋮	⋮
1.0	Enterobacter asburiae LF7a	640513

TABLE A.2 Medium Complexity Community

Rel. Ab.	Species	Taxon ID
8	Comamonas testosteroni CNB-2	688245
8	Aeromonas veronii B565	998088
9	Flavobacterium johnsoniae UW101	376686
5	Rhodoferax ferrireducens T118	338969
1.03	Cellvibrio japonicus Ueda107	498211
1.15	Lachnoclostridium phytofermentans ISDg	357809
1.16	Allochromatium vinosum DSM 180	572477
⋮	⋮	⋮
1.07	Pseudomonas fluorescens SBW25	216595

TABLE A.3 High Complexity Community

Rel. Ab.	Species	Taxon ID
1	Aeromonas veronii B565	998088
1.1	Alistipes shahii WAL 8301	717959
1.1	Acidithiobacillus ferrivorans SS3	743299
1.1	Cellulomonas flavigena DSM 20109	446466
1.0	Polynucleobacter necessarius	312153
1.1	Variovorax paradoxus EPS	595537
0.8	Herbaspirillum seropedicae SmR1	757424
⋮	⋮	⋮
1.1	Sphingomonas wittichii RW1	392499

Appendix B

Chapter 3 Supplement

B.1 Formulas and Explanations

B.1.1 Alpha diversity

Alpha diversity was measured with the `vegan` package using the Shannon (H) [129] and Simpson (D) [130] diversity indices defined as follows:

$$H = -\sum_{i=1}^{n} p_i \ln p_i$$

$$D = \frac{1}{\sum_{i=1}^{n} {p_i}^2}$$

where: n is the number of features (taxa or functions)

p_i is the proportion of reads assigned to a feature out of all features

Alpha diversity was measured using the leaves of the taxonomic tree generated from MEGAN. Since not all assignments will be made at the species level, using the leaves allows for detection of genera or less specific taxonomic groups to be factored into the diversity calculation. If only species level diversity was considered, alpha diversity might be underestimated since some taxonomic groups may only have been detected at a lower resolution and would not be factored in. An example of a simple scenario is outlined in Figure B.1. Calculation of alpha diversity was measured using the species level as well and did in fact underestimate diversity, although the same pattern was still observed. Additionally, unclassified reads were removed for alpha diversity calculations. For functional diversity, counts were used at the most specific level of SEED

classifications (level four). Only the Shannon diversity index was included in the main text, but the same pattern of diversity is seen with the Simpson index.

B.1.2 Dissimilarity Measure

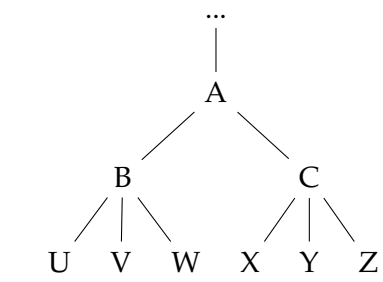
For principal coordinates analysis and multivariate dispersion, a pairwise distance matrix was generated between all sites. The Bray-Curtis dissimilarity measure [131] between samples j and sample k with n taxa was used as follows:

$$d = \frac{\sum_{i=1}^{n} |x_{ij} - x_{ik}|}{\sum_{i=1}^{n} x_{ij} + x_{ik}}$$

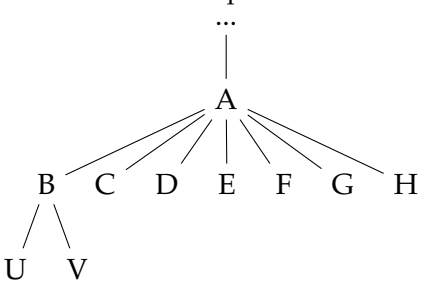
where: x is the number of read assigned to taxa i

B.1.3 Beta Diversity

Beta diversity was measured using a definition proposed by Anderson, Ellingsen, and McArdle [99]. Their definition states that the dispersion of samples around a centroid or spatial median in multivariate space can be used to determine the beta diversity of that group of samples. Beta diversity in general is the ratio of the collective diversity of a group of sites to the average diversity within each site. In a simple measure of beta diversity which examines presence-absence data, a high beta diversity would indicate that in a group of sites, the collective number of species present is much higher than the number of species detected within any one given site. If the beta diversity is low, then this means that most of the sites in the group have the same species present, and the collective grouping of species is not larger than any one site. This can also be visualized in a species accumulation plot (Figure B.2). As more sites are included in a group, the more species are detected. The greater this increase is, the greater the beta diversity. Between the sand and water samples, the slopes do not robustly suggest that there is a difference in species accumulation between the two groups of samples. This was also verified through the multivariate dispersion method, described below and reported in the main text.

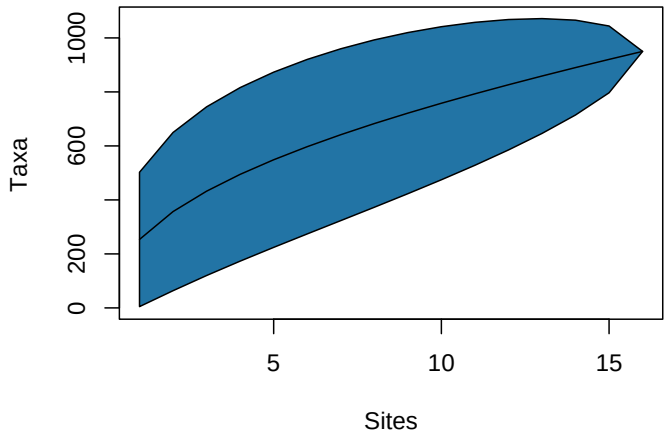


(A) Sample A

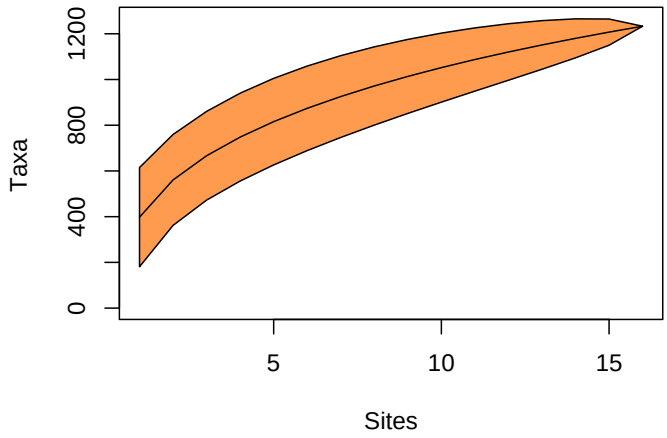


(B) Sample B

FIGURE B.1 Visualization of why taxonomic leaves were used for diversity calculations. In both samples, A is a clade at the family level. B - G are genera, and U - Z are species. In Sample A then, 6 different species are detected from two different genera. In Sample B, only 2 species are detected, but 6 other genera were detected at the genus level due to LCA. In the case where diversity calculations only look at the species level, Sample A would be seen as much more diverse since Sample B only detected two species at this clade. However, we know that each of the genera C - H detected in Sample B correspond in reality to at least one species. By using the leaves of the taxonomy tree, rather than only the species level, we can account for reads which were assigned at a non-species level, and arrive at the more realistic representation that Sample B is more diverse than Sample A. This can occur at any taxonomic level, not just the genus level (e.g. any taxa detected only at the family level would still be considered in diversity calculations).



(A) Species accumulation plot for all water samples. As more sites are introduced, the cumulative number of species detected increases. Deviations are indicated and are representative of 100 permutations.



(B) Species accumulation plot for all sand samples. As more sites are introduced, the cumulative number of species detected increases. Deviations are indicated and are representative of 100 permutations.

FIGURE B.2 Species accumulation plots for **(A)** water and **(B)** sand groups.

An expansion of the simple presence-absence measurement of beta diversity is the use of multivariate data that is not simply binary. By including relative abundance (proportional) information, the cumulative differences can be measured for a group of sites. This is what multivariate dispersion attempts to achieve. By plotting sites/samples in multivariate space, the dispersion of these sites around a centroid can serve as a measurement for how different each site is from the mean of group of sites (similar to the proportion of the number of species in all sites divided by the average number of species per site in the presence-absence example). By averaging the distance of sites to the centroid, a value is obtained which serves as the measure of beta diversity, and can be compared between groups. This is the multivariate dispersion method used in the main text for beta diversity.

Anderson, Ellingsen, and McArdle [99] also introduce a dissimilarity measure for use with their multivariate dispersion method (a modified Gower dissimilarity). However, since I could not find a study establishing the efficacy of this measurement in comparison to other dissimilarity measures, the Bray-Curtis dissimilarity measure, which has been shown to perform well in combination with proportional normalization for metagenomic data, was used [100].

B.2 Supplementary Tables

TABLE B.1 Sample Identification and Metadata

Sample Name	Identifier	Year	Beach	Lake	Site	Type	Replicate	Pair
Sample_56_BF1	1.3.2013.LS.ON	2013	LS	ON	3	Water	1	A
Sample_58_BF3	1.3.2013.LS.ON	2013	LS	ON	3	Sand	1	A
Sample_61_BF6	1.3.2013.LB.ER	2013	LB	ER	3	Water	1	B
Sample_63_BF8	1.3.2013.LB.ER	2013	LB	ER	3	Sand	1	B
Sample_62_BF7	1.5.2013.LB.ER	2013	LB	ER	5	Water	1	C
Sample_64_BF9	1.5.2013.LB.ER	2013	LB	ER	5	Sand	1	C
Sample_66_BF11	1.1.2013.LS.ON	2013	LS	ON	1	Water	1	D
Sample_67_BF12	1.1.2013.LS.ON	2013	LS	ON	1	Sand	1	D
Sample_68_BF13	1.S.2013.LB.ER	2013	LB	ER	S	Water	1	E
Sample_69_BF14	1.S.2013.LB.ER	2013	LB	ER	S	Sand	1	E
Sample_70_BF15	2.5.2013.LB.ER	2013	LB	ER	5	Water	2	F
Sample_71_BF16	2.5.2013.LB.ER	2013	LB	ER	5	Sand	2	F
Sample_72_BF17	2.3.2013.LB.ER	2013	LB	ER	3	Water	2	G
Sample_73_BF18	2.3.2013.LB.ER	2013	LB	ER	3	Sand	2	G
Sample_FP-1P	1.1.2012.FP.ON	2012	FP	ON	1	Sand	1	H
Sample_FP-1W	1.1.2012.FP.ON	2012	FP	ON	1	Water	1	H
Sample_LB-1P	1.1.2012.LB.ER	2012	LB	ER	1	Sand	1	I
Sample_LB-1W	1.1.2012.LB.ER	2012	LB	ER	1	Water	1	I
Sample_LB-3P	1.3.2012.LB.ER	2012	LB	ER	3	Sand	1	J
Sample_LB-3W	1.3.2012.LB.ER	2012	LB	ER	3	Water	1	J
Sample_LB-5P	1.5.2012.LB.ER	2012	LB	ER	5	Sand	1	K
Sample_LB-5W	1.5.2012.LB.ER	2012	LB	ER	5	Water	1	K
Sample_LCE-1P	1.C.2012.LB.ER	2012	LB	ER	C	Sand	1	L
Sample_LCE-1W	1.C.2012.LB.ER	2012	LB	ER	C	Water	1	L
Sample_LS-1P	1.1.2012.LS.ON	2012	LS	ON	1	Sand	1	M
Sample_LS-1W	1.1.2012.LS.ON	2012	LS	ON	1	Water	1	M
Sample_LS-3P	1.3.2012.LS.ON	2012	LS	ON	3	Sand	1	N
Sample_LS-3W	1.3.2012.LS.ON	2012	LS	ON	3	Water	1	N
Sample_LS-5P	1.5.2012.LS.ON	2012	LS	ON	5	Sand	1	O
Sample_LS-5W	1.5.2012.LS.ON	2012	LS	ON	5	Water	1	O
Sample_NB-1P	1.1.2012.NB.ER	2012	NB	ER	1	Sand	1	P
Sample_NB-1W	1.1.2012.NB.ER	2012	NB	ER	1	Water	1	P

B.3 Supplementary Figures

Sample Metdata												Illumina HiSeq (paired-end 2x150 bp)_Sep, 2013 Sequencing Metadata								
Sample Name		Subtype	Date Sampled	STOCK SAMPLE			SUBMITTED SAMPLE			Rnase TREATED (Y OR N)	Comments	Index	Yield (Mbases)	% PF	# Reads	% of raw clusters per lane	% Perfect Index Reads	% One Mismatch Reads (Index)	% of >= Q30 Bases (PF)	Mean Quality Score (PF)
				CONC (ng/ul)	VOLUME (ul)	TOTAL AMOUNT (ng)	CONC (ng/ul)	VOLUME (ul)	TOTAL AMOUNT (ng)											
	Lakeside 1	Water	05-Aug-12	20	20	400	1	10	10	Y										
LS-1W	Lakeside	Water	03-Aug-12	2.8	40	112	0.2	10	2	Y	Bacterial Fraction	TAGGCATG-CTAAGCCT	455	100	3014958	1	91.57	7.86	72.76	29.74
LS-3W	Lakeside	Water	03-Aug-12	2.1	40	84	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-TAGATCGC	664	100	4394854	1.46	85.68	14.01	71.39	29.29
LS-5W	Lakeside	Pore Sample	03-Aug-12	0.8	40	32	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-CTCTCTAT	522	100	3460244	1.15	92.14	7.59	71.74	29.4
LS-1P	Lakeside	Pore Sample	03-Aug-12	1.6	40	64	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-TATCCTCT	429	100	2842444	0.94	95.79	4.01	68.74	28.52
LS-3P	Lakeside	Pore Sample	03-Aug-12	12.9	40	516	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-AGAGTAGA	424	100	2810664	0.93	83.31	16.28	66.06	27.57
LS-5P	Lakeside	Pore Sample	03-Aug-12	13.5	40	540	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-GTAAGGAG	835	100	5529520	1.83	91.74	8.01	69.07	28.63
FP-1W	Fifty Point	Water	03-Aug-12	16.4	40	656	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-ACTGCATA	416	100	2757028	0.91	95.36	4.41	70.86	29.14
FP-1P	Fifty Point	Pore Sample	03-Aug-12	21.1	40	844	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-AAGGAGTA	322	100	2131564	0.71	90.14	9.5	65.9	27.64
NB-1W	Nickel Beach	Water	09-Aug-12	12.9	40	516	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-TAGATCGC	210	100	1392514	0.46	87.69	11.93	72.68	29.68
NB-1P	Nickel Beach	Pore Sample	09-Aug-12	0.24	40	9.6	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-CTCTCTAT	350	100	2320120	0.77	92.69	7.01	70.14	28.91
LB-1W	Long Beach	Water	09-Aug-12	2.4	40	96	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-TATCCTCT	652	100	4320828	1.43	96.32	3.49	72.09	29.54
LB-3W	Long Beach	Water	09-Aug-12	0.9	40	36	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-AGAGTAGA	524	100	3473310	1.15	84.99	14.6	64.66	27.1
LB-5W	Long Beach	Water	09-Aug-12	6.4	40	256	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-GTAAGGAG	538	100	3560266	1.18	92.02	7.69	67.14	28
LB-1P	Long Beach	Pore Sample	09-Aug-12	2.3	40	92	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-ACTGCATA	480	100	3177520	1.05	95.36	4.39	67.84	28.2
LB-3P	Long Beach	Pore Sample	09-Aug-12	4.2	40	168	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-AAGGAGTA	616	100	4079142	1.35	91.22	8.46	67.85	28.21
LB-5P	Long Beach	Pore Sample	09-Aug-12	6.9	40	276	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-CTAAGCCT	478	100	3166998	1.05	92.72	7	69.37	28.7
LCE-1W	Long Beach C.E.	Water	09-Aug-12	1.2	40	48	0.2	10	2	Y	Bacterial Fraction	GCTACGCT-TAGATCGC	461	100	3051230	1.01	88.27	11.31	74.56	30.31
LCE-1P	Long Beach C.E.	Pore Sample	09-Aug-12	35.7	40	1428	0.2	10	2	Y	Bacterial Fraction	GCTACGCT-CTCTCTAT	542	100	3590974	1.19	92.75	6.87	71.22	29.31
												Illumina HiSeq (paired-end 2x100 bp)_Jul, 2014								
BF1	LS-3W	Water	2013-07-31	1.748		69.92	0.2	10	2	Y	Bacterial Fraction	TAAGGCGA-ACTGCATA	188	100	1864734	0.72	95.49	4.25	79.27	31.6
BF3	LS-3P	Pore Sample	2013-07-31	43.8		1752	0.2	10	2	Y	Bacterial Fraction	AGGCAGAA-ACTGCATA	108	100	1072990	0.42	95.89	3.85	77.03	30.81
BF6	LB-3W	Water	2013-08-14	4.18		167.2	0.2	10	2	Y	Bacterial Fraction	TAGGCATG-ACTGCATA	159	100	1575264	0.61	93.37	6.42	82.22	32.48
BF7	LB-5W	Water	2013-08-14	2.08		83.2	0.2	10	2	Y	Bacterial Fraction	CTCTCTAC-ACTGCATA	215	100	2129238	0.82	96.95	2.82	83.26	32.75
BF8	LB-3P	Pore Sample	2013-08-14	9.04		361.6	0.2	10	2	Y	Bacterial Fraction	CAGAGAGG-ACTGCATA	165	100	1635940	0.63	96.84	2.94	78.33	31.23
BF9	LB-5P	Pore Sample	2013-08-14	12.36		494.4	0.2	10	2	Y	Bacterial Fraction	GCTACGCT-ACTGCATA	203	100	2005902	0.78	96.6	3.22	85.11	33.44
BF11	LS-1W	Water	2013-08-21	3.16		126.4	0.2	10	2	Y	Bacterial Fraction	AAGAGGCA-ACTGCATA	315	100	3114004	1.21	97.25	2.59	85.52	33.58
BF12	LS-1P	Pore Sample	2013-08-21	25.4		1016	0.2	10	2	Y	Bacterial Fraction	GTAGAGGA-ACTGCATA	241	100	2385936	0.92	94.93	4.82	83.85	32.96
BF13	LBS-1W	Water	2013-08-28	1.852		74.08	0.2	10	2	Y	Bacterial Fraction	TAAGGCGA-AAGGAGTA	155	100	1536274	0.59	94.17	5.34	79.11	31.58
BF14	LBS-1P	Source-Pore	2013-08-28	9.18		367.2	0.2	10	2	Y	Bacterial Fraction	CGTACTAG-AAGGAGTA	221	100	2191774	0.85	95.75	4.03	81.16	32.11
BF15	LB-5W	Water	2013-08-28	7.3		292	0.2	10	2	Y	Bacterial Fraction	AGGCAGAA-AAGGAGTA	144	100	1430482	0.55	94.68	4.89	78.53	31.36
BF16	LB-5P	Pore Sample	2013-08-28	6.42		256.8	0.2	10	2	Y	Bacterial Fraction	TCCTGAGC-AAGGAGTA	193	100	1910316	0.74	95.28	4.35	82.55	32.68
BF17	LB-3W	Water	2013-08-28	3.02		120.8	0.2	10	2	Y	Bacterial Fraction	GGACTCCT-AAGGAGTA	170	100	1686544	0.65	95.6	4.09	85.56	33.55
BF18	LB-3P	Pore Sample	2013-08-28	4.36		174.4	0.2	10	2	Y	Bacterial Fraction	TAGGCATG-AAGGAGTA	172	100	1705220	0.66	92.65	6.9	83.55	32.89

FIGURE B.3 Metadata and sequencing information of samples analyzed.

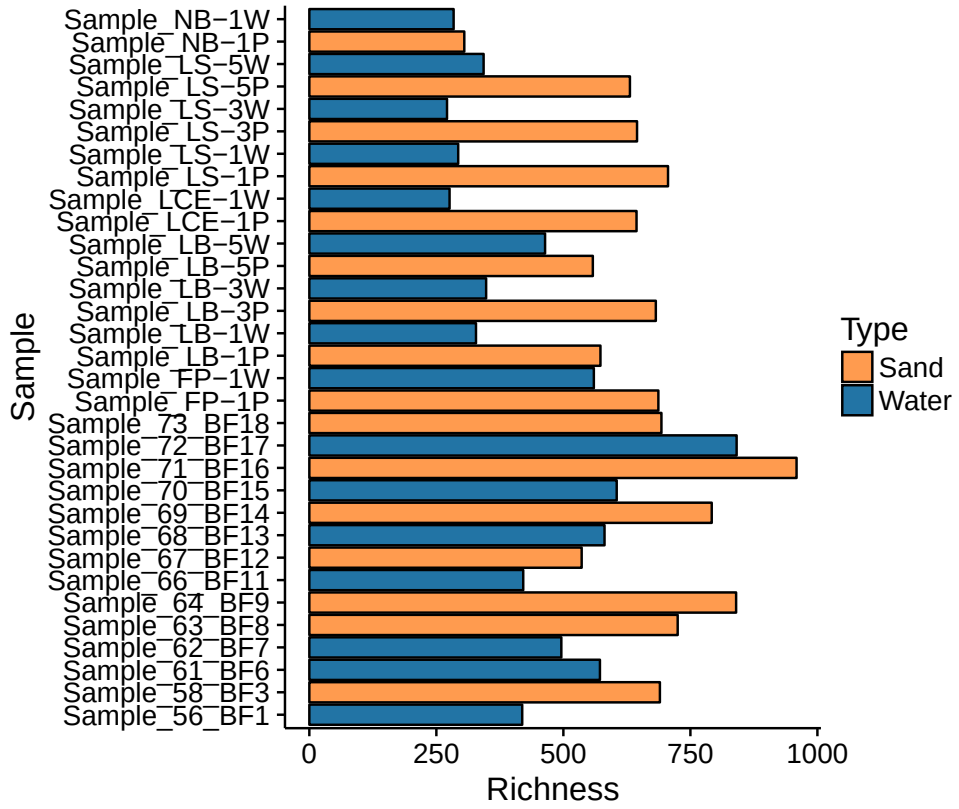


FIGURE B.4 Taxonomic richness across all sites examined. Richness was measured as the number of unique taxonomic features detected.

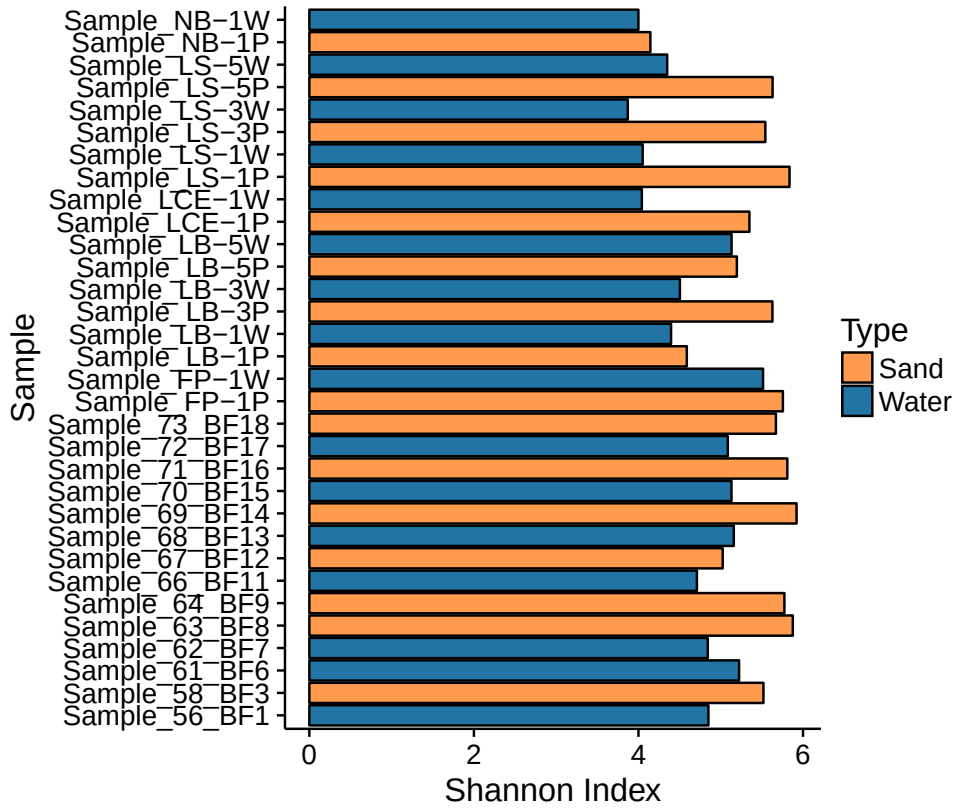
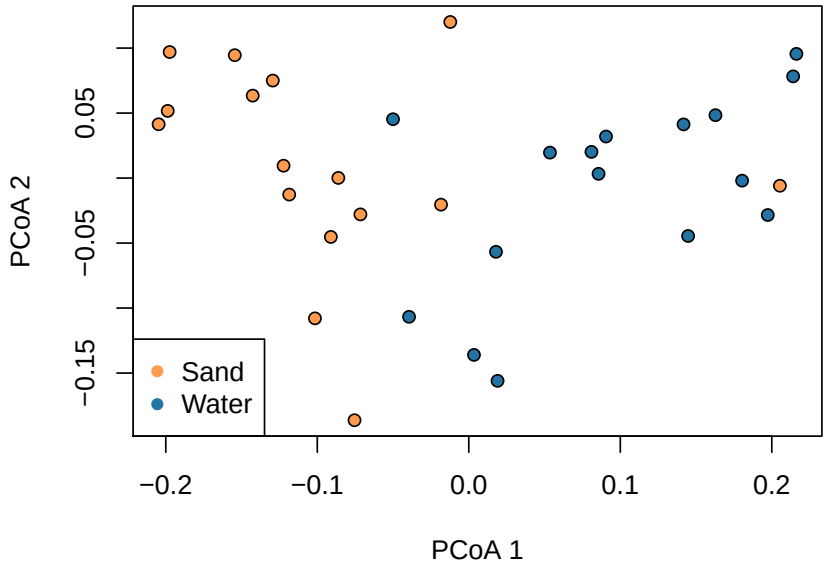
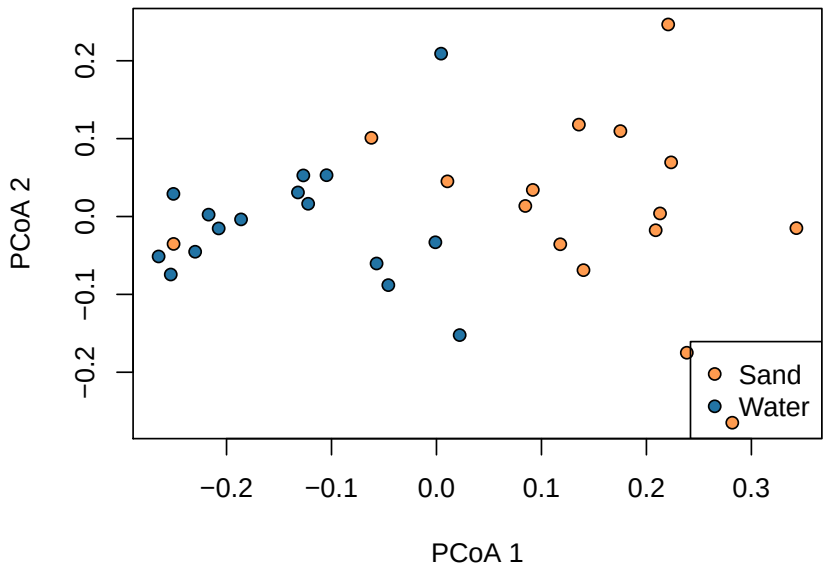


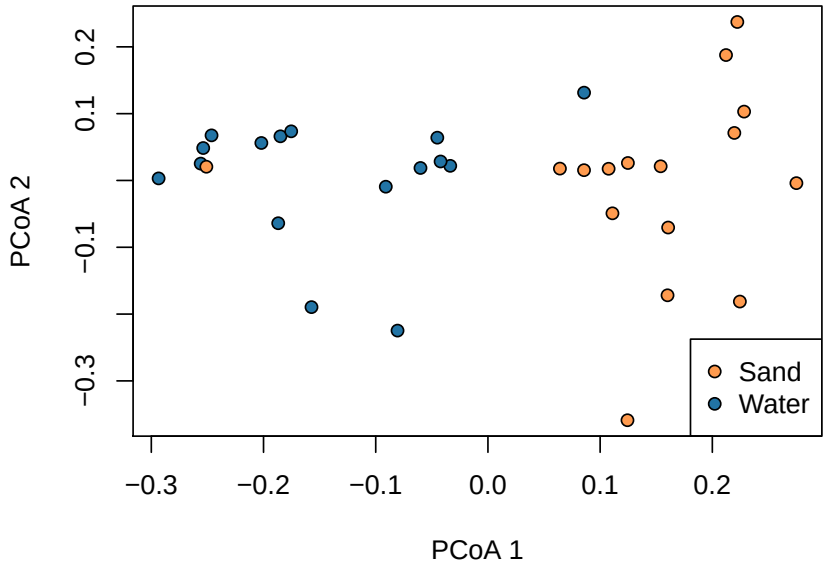
FIGURE B.5 Taxonomic diversity across all sites examined. Taxonomic diversity was measured using the Shannon index.



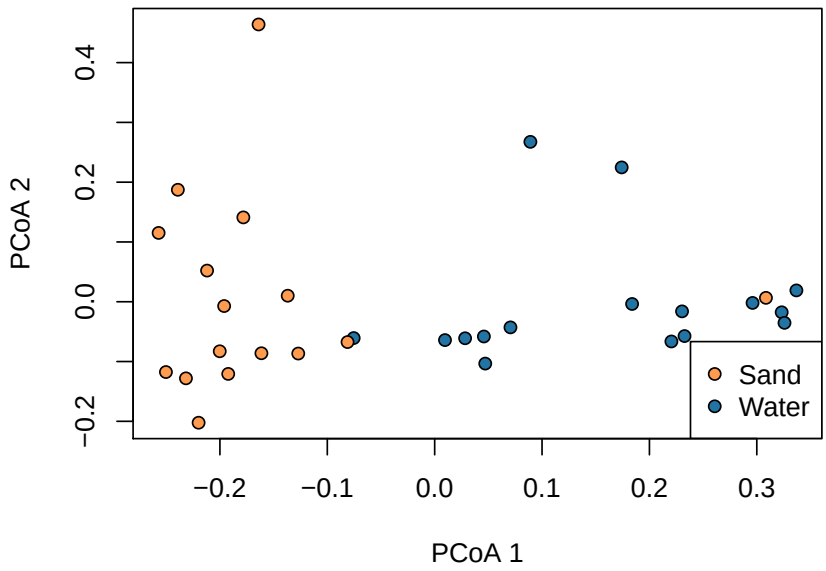
(A) Phylum



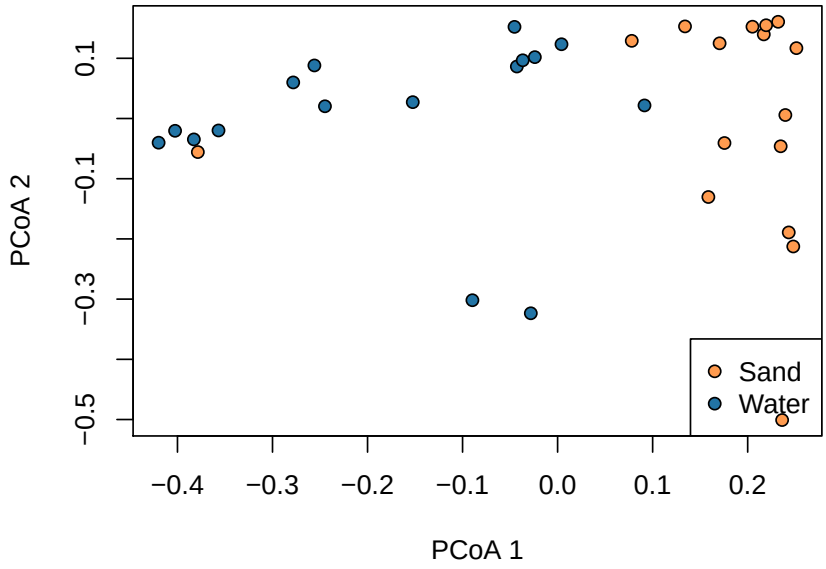
(B) Class



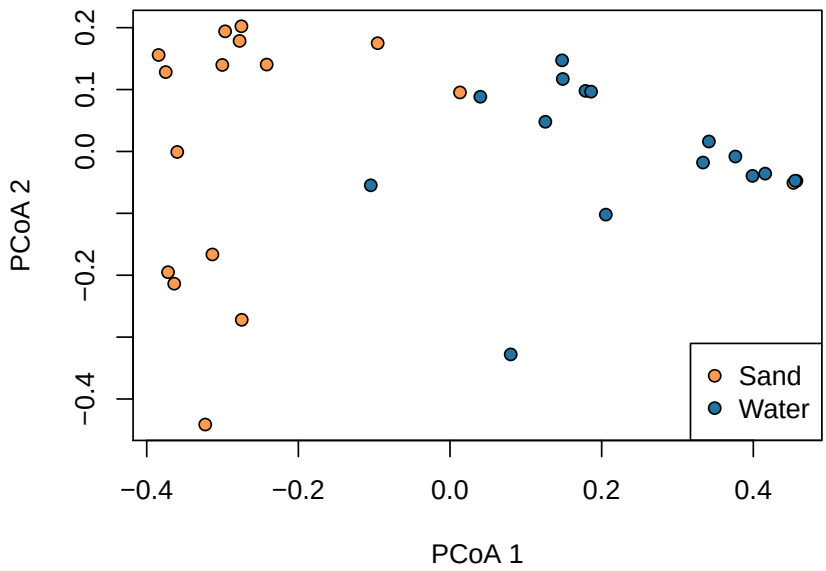
(C) Order



(D) Family



(E) Genus



(F) Species

FIGURE B.6 Principal coordinates analysis of Bray-Curtis dissimilarities at varying taxonomic ranks.

Bibliography

[1] W.B. Whitman, D.C. Coleman, and W.J. Wiebe. 1998. "Prokaryotes: the unseen majority". *Proc Natl Acad Sci U S A* 95, pp. 6578–6583.

[2] G. Muyzer, A. Teske, C.O. Wirsen, and H.W. Jannasch. 1995. "Phylogenetic relationships of Thiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments". *Arch Microbiol* 164, pp. 165–172.

[3] V.J. Orphan, K.U. Hinrichs, W.3 Ussler, C.K. Paull, L.T. Taylor, S.P. Sylva, J.M. Hayes, and E.F. Delong. 2001. "Comparative analysis of methane-oxidizing archaea and sulfate-reducing bacteria in anoxic marine sediments". *Appl Environ Microbiol* 67, pp. 1922–1934. DOI: 10.1128/AEM.67.4.1922-1934.2001.

[4] K.B. Hallberg and D.B. Johnson. 2003. "Novel acidophiles isolated from moderately acidic mine drainage waters". *Hydrometallurgy* 71.1, pp. 139–148.

[5] G.W. O'Hara. 2001. "Nutritional constraints on root nodule bacteria affecting symbiotic nitrogen fixation: a review". *Animal Production Science* 41.3, pp. 417–433.

[6] P.J. Turnbaugh, R.E. Ley, M.A. Mahowald, V. Magrini, E.R. Mardis, and J.I. Gordon. 2006. "An obesity-associated gut microbiome with increased capacity for energy harvest". *Nature* 444, pp. 1027–1031. DOI: 10.1038/nature05414.

[7] F. Backhed, H. Ding, T. Wang, L.V. Hooper, G.Y. Koh, A. Nagy, C.F. Semenkovich, and J.I. Gordon. 2004. "The gut microbiota as an environmental factor that regulates fat storage". *Proc Natl Acad Sci U S A* 101, pp. 15718–15723. DOI: 10.1073/pnas.0407076101.

[8] P.J. Turnbaugh, R.E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J.I. Gordon. 2007. "The human microbiome project: exploring the microbial part of ourselves in a changing world". *Nature* 449.7164, p. 804.

[9] F. vonNussbaum, M. Brands, B. Hinzen, S. Weigand, and D. Habich. 2006. "Antibacterial natural products in medicinal chemistry–exodus or revival?" *Angew Chem Int Ed Engl* 45, pp. 5072–5129. DOI: 10.1002/anie.200600350.

[10] J. Handelsman. 2004. "Metagenomics: application of genomics to uncultured microorganisms". *Microbiol Mol Biol Rev* 68, pp. 669–685. DOI: 10.1128/MMBR.68.4.669-685.2004.

[11] N.R. Pace, D.A. Stahl, D.J. Lane, and G.J. Olsen. 1986. "The analysis of natural microbial populations by ribosomal RNA sequences". *Advances in microbial ecology*. Springer, pp. 1–55.

[12] J.T. Staley and A. Konopka. 1985. "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats". *Annu Rev Microbiol* 39, pp. 321–346. DOI: 10.1146/annurev.mi.39.100185.001541.

[13] V. Torsvik, J. Goksoyr, and F.L. Daae. 1990. "High diversity in DNA of soil bacteria". *Appl Environ Microbiol* 56, pp. 782–787.

[14] J.G. Jones. 1977. "The effect of environmental factors on estimated viable and total populations of planktonic bacteria in lakes and experimental enclosures". *Freshwater Biol* 7.1, pp. 67–91.

[15] K. Kogure, U. Simidu, and N. Taga. 1979. "A tentative direct microscopic method for counting living marine bacteria". *Can J Microbiol* 25, pp. 415–420.

[16] R.I. Amann, W. Ludwig, and K.H. Schleifer. 1995. "Phylogenetic identification and in situ detection of individual microbial cells without cultivation". *Microbiol Rev* 59, pp. 143–169.

[17] D.A. Stahl, D.J. Lane, G.J. Olsen, and N.R. Pace. 1984. "Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences". *Science* 224, pp. 409–411. DOI: 10.1126/science.224.4647.409.

[18] — 1985. "Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences". *Appl Environ Microbiol* 49, pp. 1379–1384.

[19] T.M. Schmidt, E.F. DeLong, and N.R. Pace. 1991. "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing". *J Bacteriol* 173, pp. 4371–4378.

[20] S.J. Giovannoni, T.B. Britschgi, C.L. Moyer, and K.G. Field. 1990. "Genetic diversity in Sargasso Sea bacterioplankton". *Nature* 345, pp. 60–63. DOI: 10.1038/345060a0.

[21] D.M. Ward, R. Weller, and M.M. Bateson. 1990. "16S rRNA sequences reveal numerous uncultured microorganisms in a natural community". *Nature* 345, pp. 63–65. DOI: 10.1038/345063a0.

[22] J.R. Marchesi, T. Sato, A.J. Weightman, T.A. Martin, J.C. Fry, S.J. Hiom, D. Dymock, and W.G. Wade. 1998. "Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA". *Appl Environ Microbiol* 64, pp. 795–799.

[23] G.C. Baker, J.J. Smith, and D.A. Cowan. 2003. "Review and re-analysis of domain-specific 16S primers". *J Microbiol Methods* 55, pp. 541–555.

[24] S. Kittelmann, H. Seedorf, W.A. Walters, J.C. Clemente, R. Knight, J.I. Gordon, and P.H. Janssen. 2013. "Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities". *PLoS One* 8, e47879. DOI: 10.1371/journal.pone.0047879.

[25] M.I. Uyaguari-Diaz, M. Chan, B.L. Chaban, M.A. Croxen, J.F. Finke, J.E. Hill, M.A. Peabody, T. VanRossum, C.A. Suttle, F.S. Brinkman, J. Isaac-Renton, N.A. Prystajecky, and P. Tang. 2016. "A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples". *Microbiome* 4, p. 20. DOI: 10.1186/s40168-016-0166-1.

[26] N.R. Pace. 2009. "Mapping the tree of life: progress and prospects". *Microbiol Mol Biol Rev* 73, pp. 565–576. DOI: 10.1128/MMBR.00033-09.

[27] J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.H. Rogers, and H.O. Smith. 2004. "Environmental genome shotgun sequencing of the Sargasso Sea". *Science* 304, pp. 66–74. DOI: 10.1126/science.1093857.

[28] G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. "Community structure and metabolism through reconstruction of microbial genomes from the environment". *Nature* 428, pp. 37–43. DOI: 10.1038/nature02340.

[29] S.G. Tringe, C. vonMering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. "Comparative metagenomics of microbial communities". *Science* 308, pp. 554–557. DOI: 10.1126/science.1107851.

[30] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. "Genome sequencing in microfabricated high-density picolitre reactors". *Nature* 437, pp. 376–380. DOI: 10.1038/nature03959.

[31] D.R. Bentley et al. 2008. "Accurate whole human genome sequencing using reversible terminator chemistry". *Nature* 456, pp. 53–59. DOI: 10.1038/nature07517.

[32] R. Logares, T.H. Haverkamp, S. Kumar, A. Lanzen, A.J. Nederbragt, C. Quince, and H. Kauserud. 2012. "Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches". *J Microbiol Methods* 91, pp. 106–113. DOI: 10.1016/j.mimet.2012.07.017.

[33] C. Pedros-Alio. 2006. "Marine microbial diversity: can it be determined?" *Trends Microbiol* 14, pp. 257–263. DOI: 10.1016/j.tim.2006.04.007.

[34] M.B. Scholz, C.C. Lo, and P.S. Chain. 2012. "Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis". *Curr Opin Biotechnol* 23, pp. 9–15. DOI: 10.1016/j.copbio.2011.11.013.

[35] Z. Liu, C. Lozupone, M. Hamady, F.D. Bushman, and R. Knight. 2007. "Short pyrosequencing reads suffice for accurate microbial community analysis". *Nucleic Acids Res* 35, e120. DOI: 10.1093/nar/gkm541.

[36] R. Ranjan, A. Rani, A. Metwally, H.S. McGee, and D.L. Perkins. 2016. "Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing". *Biochem Biophys Res Commun* 469, pp. 967–977. DOI: 10.1016/j.bbrc.2015.12.083.

[37] A. Benitez-Paez, K.J. Portune, and Y. Sanz. 2016. "Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION portable nanopore sequencer". *Gigascience* 5, p. 4. DOI: 10.1186/s13742-016-0111-z.

[38] E.A. Eloe-Fadrosh, N.N Ivanova, T. Woyke, and N.C. Kyrpides. 2016. "Metagenomics uncovers gaps in amplicon-based detection of microbial diversity". *Nat Microbiol* 1.

[39] C.T. Brown, L.A. Hug, B.C. Thomas, I. Sharon, C.J. Castelle, A. Singh, M.J. Wilkins, K.C. Wrighton, K.H. Williams, and J.F. Banfield. 2015. "Unusual biology across a group comprising more than 15% of domain Bacteria". *Nature* 523, pp. 208–211. DOI: 10.1038/nature14486.

[40] J.L. Morgan, A.E. Darling, and J.A. Eisen. 2010. "Metagenomic sequencing of an in vitro-simulated microbial community". *PLoS One* 5, e10209. DOI: 10.1371/journal.pone.0010209.

[41] N. Shah, H. Tang, T.G. Doak, and Y. Ye. 2011b. "Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics", pp. 165–176.

[42] H. Teeling and F.O. Glockner. 2012. "Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective". *Brief Bioinform* 13, pp. 728–742. DOI: 10.1093/bib/bbs039.

[43] B.J. Finlay. 2002. "Global dispersal of free-living microbial eukaryote species". *Science* 296, pp. 1061–1063. DOI: 10.1126/science.1070710.

[44] L.Y. Stein, M.T. LaDuc, T.J. Grundl, and K.H. Nealson. 2001. "Bacterial and archaeal populations associated with freshwater ferromanganous micronodules and sediments". *Environ Microbiol* 3, pp. 10–18.

[45] C.S. Lee, C. Lee, J. Marion, Q. Wang, L. Saif, and J. Lee. 2014. "Occurrence of human enteric viruses at freshwater beaches during swimming season and its link to water inflow". *Sci Total Environ* 472, pp. 757–766. DOI: 10.1016/j.scitotenv.2013.11.088.

[46] M. Mohiuddin and H.E. Schellhorn. 2015. "Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis". *Front Microbiol* 6, p. 960. DOI: 10.3389/fmicb.2015.00960.

[47] S. Oh, A. Caro-Quintero, D. Tsementzi, N. DeLeon-Rodriguez, C. Luo, R. Poretsky, and K.T. Konstantinidis. 2011. "Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem". *Appl Environ Microbiol* 77, pp. 6000–6011. DOI: 10.1128/AEM.00107-11.

[48] M.M. Steffen, Z. Li, T.C. Effler, L.J. Hauser, G.L. Boyer, and S.W. Wilhelm. 2012. "Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents". *PLoS One* 7, e44002. DOI: 10.1371/journal.pone.0044002.

[49] A.K. Sharma, K. Sommerfeld, G.S. Bullerjahn, A.R. Matteson, S.W. Wilhelm, J. Jezbera, U. Brandt, W.F. Doolittle, and M.W. Hahn. 2009. "Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria". *ISME J* 3, pp. 726–737. DOI: 10.1038/ismej.2009.13.

[50] S.L. Garcia, K.D. McMahon, M. Martinez-Garcia, A. Srivastava, A. Sczyrba, R. Stepanauskas, H.P. Grossart, T. Woyke, and F. Warnecke. 2013. "Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton". *ISME J* 7, pp. 137–147. DOI: 10.1038/ismej.2012.86.

[51] X. Mou, X. Lu, J. Jacob, S. Sun, and R. Heath. 2013. "Metagenomic identification of bacterioplankton taxa and pathways involved in microcystin degradation in lake erie". *PLoS One* 8, e61890. DOI: 10.1371/journal.pone.0061890.

[52] S.S. Mande, M.H. Mohammed, and T.S. Ghosh. 2012. "Classification of metagenomic sequences: methods and challenges". *Brief Bioinform* 13, pp. 669–681. DOI: 10.1093/bib/bbs054.

[53] L. Pasic, B. Rodriguez-Mueller, A.B. Martin-Cuadrado, A. Mira, F. Rohwer, and F. Rodriguez-Valera. 2009. "Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*". *BMC Genomics* 10, p. 570. DOI: 10.1186/1471-2164-10-570.

[54] D.H. Huson, A.F. Auch, J. Qi, and S.C. Schuster. 2007. "MEGAN analysis of metagenomic data". *Genome Res* 17, pp. 377–386. DOI: 10.1101/gr.5969107.

[55] S. Karlin, J. Mrazek, and A.M. Campbell. 1997. "Compositional biases of bacterial genomes and evolutionary implications". *J Bacteriol* 179, pp. 3899–3913.

[56] A.C. McHardy, H.G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. 2007. "Accurate phylogenetic classification of variable-length DNA fragments". *Nat Methods* 4, pp. 63–72. DOI: 10.1038/nmeth976.

[57] G.L. Rosen, E.R. Reichenberger, and A.M. Rosenfeld. 2011. "NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads". *Bioinformatics* 27, pp. 127–129. DOI: 10.1093/bioinformatics/btq619.

[58] A. Brady and S.L. Salzberg. 2009. "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models". *Nat Methods* 6, pp. 673–676. DOI: 10.1038/nmeth.1358.

[59] D.E. Wood and S.L. Salzberg. 2014. "Kraken: ultrafast metagenomic sequence classification using exact alignments". *Genome Biol* 15, R46. DOI: 10.1186/gb-2014-15-3-r46.

[60] S.K. Ames, D.A. Hysom, S.N. Gardner, G.S. Lloyd, M.B. Gokhale, and J.E. Allen. 2013. "Scalable metagenomic taxonomy classification using a reference genome database". *Bioinformatics* 29, pp. 2253–2260. DOI: 10.1093/bioinformatics/btt389.

[61] A.L. Bazinet and M.P. Cummings. 2012. "A comparative evaluation of sequence classification programs". *BMC Bioinformatics* 13, p. 92. DOI: 10.1186/1471-2105-13-92.

[62] J. Droge, I. Gregor, and A.C. McHardy. 2015. "Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods". *Bioinformatics* 31, pp. 817–824. DOI: 10.1093/bioinformatics/btu745.

[63] R. Ounit, S. Wanamaker, T.J. Close, and S. Lonardi. 2015. "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers". *BMC Genomics* 16, p. 236. DOI: 10.1186/s12864-015-1419-2.

[64] A. Charuvaka and H. Rangwala. 2011. "Evaluation of short read metagenomic assembly". *BMC Genomics* 12 Suppl 2, S8. DOI: 10.1186/1471-2164-12-S2-S8.

[65] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. "Basic local alignment search tool". *J Mol Biol* 215, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[66] B. Buchfink, C. Xie, and D.H. Huson. 2015. "Fast and sensitive protein alignment using DIAMOND". *Nat Methods* 12, pp. 59–60. DOI: 10.1038/nmeth.3176.

[67] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, and H. Nielsen. 2000. "Assessing the accuracy of prediction algorithms for classification: an overview". *Bioinformatics* 16, pp. 412–424.

[68] N.N. Diaz, L. Krause, A. Goesmann, K. Niehaus, and T.W. Nattkemper. 2009. "TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach". *BMC Bioinformatics* 10, p. 56. DOI: 10.1186/1471-2105-10-56.

[69] M.A. Peabody, T. VanRossum, R. Lo, and F.S. Brinkman. 2015. "Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities". *BMC Bioinformatics* 16, p. 363. DOI: 10.1186/s12859-015-0788-5.

[70] S. Lindgreen, K.L. Adair, and P.P. Gardner. 2016. "An evaluation of the accuracy and speed of metagenome analysis tools". *Sci Rep* 6, p. 19233. DOI: 10.1038/srep19233.

[71] Environment and Climate Change Canada. 2015. *Fresh Water Quality Monitoring and Surveillance*. URL: http://www.ec.gc.ca/eaudouce-freshwater/ (visited on 06/19/2016).

[72] USGS. 2016. *Water-Quality Methods and Techniques*. URL: http://water.usgs.gov/owq/methods.html (visited on 06/19/2016).

[73] H. Shuval. 2003. "Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment". *J Water Health* 1, pp. 53–64.

[74] S. Sharma, P. Sachdeva, and J.S. Virdi. 2003. "Emerging water-borne pathogens". *Appl Microbiol Biotechnol* 61, pp. 424–428. DOI: 10.1007/s00253-003-1302-y.

[75] M. Leblanc-Maridor, F. Beaudeau, H. Seegers, M. Denis, and C. Belloc. 2011. "Rapid identification and quantification of Campylobacter coli and Campylobacter jejuni by real-time PCR in pure cultures and in complex samples". *BMC Microbiol* 11, p. 113. DOI: 10.1186/1471-2180-11-113.

[76] T.C. Lee, R.M. Vickers, V.L. Yu, and M.M. Wagener. 1993. "Growth of 28 *Legionella* species on selective culture media: a comparative study". *J Clin Microbiol* 31, pp. 2764–2768.

[77] E. Halliday and R.J. Gast. 2011. "Bacteria in beach sands: an emerging challenge in protecting coastal water quality and bather health". *Environ Sci Technol* 45, pp. 370–379. DOI: 10.1021/es102747s.

[78] T.J. Wade, R.L. Calderon, E. Sams, M. Beach, K.P. Brenner, A.H. Williams, and A.P. Dufour. 2006. "Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness". *Environ Health Perspect* 114, pp. 24–28.

[79] D.N. Myers, D.M. Stoeckel, R.N. Bushon, D.S. Francy, and A.M. Brady. 2014. "Fecal Indicator Bacteria". *U.S. Geological Survey Techniques of Water-Resources Investigations*. Chap. A7-7.1.

[80] J.W. Marion, J. Lee, S. Lemeshow, and T.J. Buckley. 2010. "Association of gastrointestinal illness and recreational water exposure at an inland U.S. beach". *Water Res* 44, pp. 4796–4804. DOI: 10.1016/j.watres.2010.07.065.

[81] J.r. ColfordJM, T.J. Wade, K.C. Schiff, C.C. Wright, J.F. Griffith, S.K. Sandhu, S. Burns, M. Sobsey, G. Lovelace, and S.B. Weisberg. 2007. "Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination". *Epidemiology* 18, pp. 27–35. DOI: 10.1097/01.ede.0000249425.32990.b9.

[82] T.J. Wade, N. Pai, J.N. Eisenberg, and J.r. ColfordJM. 2003. "Do U.S. Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis". *Environ Health Perspect* 111, pp. 1102–1109.

[83] T.M. Straub and D.P. Chandler. 2003. "Towards a unified system for detecting waterborne pathogens". *J Microbiol Methods* 53, pp. 185–197.

[84] R.L. Whitman and M.B. Nevers. 2003. "Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach". *Appl Environ Microbiol* 69, pp. 5555–5562.

[85] M.N. Byappanahalli, R.L. Whitman, D.A. Shively, W.T. Ting, C.C. Tseng, and M.B. Nevers. 2006. "Seasonal persistence and population characteristics of *Escherichia coli* and enterococci in deep backshore sand of two freshwater beaches". *J Water Health* 4, pp. 313–320.

[86] A. Hartz, M. Cuvelier, K. Nowosielski, T.D. Bonilla, M. Green, N. Esiobu, D.S. McCorquodale, and A. Rogerson. 2008. "Survival potential of Escherichia coli and Enterococci in subtropical beach sand: implications for water quality managers". *J Environ Qual* 37, pp. 898–905. DOI: 10.2134/jeq2007.0312.

[87] E. Wheeler-Alm, J. Burke, and A. Spain. 2003. "Fecal indicator bacteria are abundant in wet sand at freshwater beaches". *Water Res* 37, pp. 3978–3982. DOI: 10.1016/S0043-1354(03)00301-4.

[88] S. Ishii, D. L. Hansen, R. E. Hicks, and M. J. Sadowsky. 2007. "Beach sand and sediments are temporal sinks and sources of *Escherichia coli* in Lake Superior". *Environmental science & technology* 41.7, pp. 2203–2209.

[89] I.U. Khan, A. Loughborough, and T.A. Edge. 2009. "DNA-based real-time detection and quantification of aeromonads from fresh water beaches on Lake Ontario". *J Water Health* 7, pp. 312–323. DOI: 10.2166/wh.2009.041.

[90] K.M. Yamahara, L.M. Sassoubre, K.D. Goodwin, and A.B. Boehm. 2012. "Occurrence and persistence of bacterial pathogens and indicator organisms in beach sand along the California coast". *Appl Environ Microbiol* 78, pp. 1733–1745. DOI: 10.1128/AEM.06185-11.

[91] F.J. Bolton, S.B. Surman, K. Martin, D.R. Wareing, and T.J. Humphrey. 1999. "Presence of Campylobacter and Salmonella in sand from bathing beaches". *Epidemiol Infect* 122, pp. 7–13.

[92] A.H. Shah, A.M. Abdelzaher, M. Phillips, R. Hernandez, H.M. Solo-Gabriele, J. Kish, G. Scorzetti, J.W. Fell, M.R. Diaz, T.M. Scott, J. Lukasik, V.J. Harwood, S. McQuaig, C.D. Sinigalliano, M.L. Gidley, D. Wanless, A. Ager, J. Lui, J.R. Stewart, L.R. Plano, and L.E. Fleming. 2011a. "Indicator microbes correlate with pathogenic bacteria, yeasts and helminthes in sand at a subtropical recreational beach site". *J Appl Microbiol* 110, pp. 1571–1583. DOI: 10.1111/j.1365-2672.2011.05013.x.

[93] Niagara Region. 2016. *Beach Water Testing in Niagara*. URL: http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm (visited on 06/17/2016).

[94] Q. Zhou, X. Su, and K. Ning. 2014. "Assessment of quality control approaches for metagenomic data analysis". *Sci Rep* 4, p. 6957. DOI: 10.1038/srep06957.

[95] E.W. Sayers, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. 2009. "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Res* 37, pp. D5–15. DOI: 10.1093/nar/gkn741.

[96] R. Overbeek, R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A.R. Wattam, F. Xia, and R. Stevens. 2014. "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)". *Nucleic Acids Res* 42, pp. D206–14. DOI: 10.1093/nar/gkt1226.

[97] D. Simberloff. 1978. "Use of rarefaction and related methods in ecology". *Biological data in water pollution assessment: quantitative and statistical analyses*. ASTM International.

[98] P. Koleff, K. J. Gaston, and J. J. Lennon. 2003. "Measuring beta diversity for presence–absence data". *J Anim Ecol* 72.3, pp. 367–382.

[99] M.J. Anderson, K.E. Ellingsen, and B.H. McArdle. 2006. "Multivariate dispersion as a measure of beta diversity". *Ecol Lett* 9, pp. 683–693. DOI: 10.1111/j.1461-0248.2006.00926.x.

[100] P.J. McMurdie and S. Holmes. 2014. "Waste not, want not: why rarefying microbiome data is inadmissible". *PLoS Comput Biol* 10, e1003531. DOI: 10.1371/journal.pcbi.1003531.

[101] M. J. Anderson. 2001. "A new method for non-parametric multivariate analysis of variance". *Austral Ecology* 26.1, pp. 32–46.

[102] M. Anderson and D. Walsh. 2013. "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing?" *Ecological Monographs* 83.4, pp. 557–574.

[103] R.S. Gupta, M. Pereira, C. Chandrasekera, and V. Johari. 2003. "Molecular signatures in protein sequences that are characteristic of cyanobacteria and plastid homologues". *Int J Syst Evol Microbiol* 53, pp. 1833–1842. DOI: 10.1099/ijs.0.02720-0.

[104] W. Hou, S. Wang, H. Dong, H. Jiang, B.R. Briggs, J.P. Peacock, Q. Huang, L. Huang, G. Wu, X. Zhi, W. Li, J.A. Dodsworth, B.P. Hedlund, C. Zhang, H.E. Hartnett, P. Dijkstra, and B.A. Hungate. 2013. "A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing". *PLoS One* 8, e53350. DOI: 10.1371/journal.pone.0053350.

[105] K.F. Ettwig, M.K. Butler, D. Le Paslier, E. Pelletier, S. Mangenot, M. Kuypers, F. Schreiber, B.E. Dutilh, J. Zedelius, D. De Beer, et al. 2010. "Nitrite-driven anaerobic methane oxidation by oxygenic bacteria". *Nature* 464.7288, pp. 543–548.

[106] N.H. Youssef, I.F. Farag, C. Rinke, S.J. Hallam, T. Woyke, and M.S. Elshahed. 2015. "In Silico Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum "Latescibacteria" (WS3)". *PLoS One* 10, e0127499. DOI: 10.1371/journal.pone.0127499.

[107] S.J. Weiss, Z. Xu, A. Amir, S. Peddada, K. Bittinger, A. Gonzalez, C. Lozupone, J.R. Zaneveld, Y. Vazquez-Baeza, A. Birmingham, and R. Knight. 2015. *Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data*. Tech. rep. PeerJ PrePrints. Pre-published.

[108] R. Ghai, C.M. Mizuno, A. Picazo, A. Camacho, and F. Rodriguez-Valera. 2014. "Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing". *Mol Ecol* 23, pp. 6073–6090. DOI: 10.1111/mec.12985.

[109] R.J. Newton, S.E. Jones, A. Eiler, K.D. McMahon, and S. Bertilsson. 2011. "A guide to the natural history of freshwater lake bacteria". *Microbiol Mol Biol Rev* 75, pp. 14–49. DOI: 10.1128/MMBR.00028-10.

[110] Y. Sekiguchi, A. Ohashi, D.H. Parks, T. Yamauchi, G.W. Tyson, and P. Hugenholtz. 2015. "First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking". *PeerJ* 3, e740. DOI: 10.7717/peerj.740.

[111] Y. Jiang, G. Xu, and H. Xu. 2016. "Use of multivariate dispersion to assess water quality based on species composition data". *Environ Sci Pollut Res Int* 23, pp. 3267–3272. DOI: 10.1007/s11356-015-5583-3.

[112] R. Whitman, V.J. Harwood, T.A. Edge, M. Nevers, M. Byappanahalli, K. Vijayavel, J. Brandao, M.J. Sadowsky, E.W. Alm, A. Crowe, D. Ferguson, Z. Ge,

E. Halliday, J. Kinzelman, G. Kleinheinz, K. Przybyla-Kelly, C. Staley, Z. Staley, and H.M. Solo-Gabriele. 2014. "Microbes in Beach Sands: Integrating Environment, Ecology and Public Health". 13, pp. 329–368. DOI: 10.1007/s11157-014-9340-8.

[113] C.S.W Kueh, T. Tam, T. Lee, S.L. Wong, O.L. Lloyd, I.T.S. Yu, T.W. Wong, J.S. Tam, and D.C.J Bassett. 1995. "Epidemiological study of swimming-associated illnesses relating to bathing-beach water quality". *Water Science and Technology* 31.5, pp. 1–4.

[114] A.E. Santoro and A.B. Boehm. 2007. "Frequent occurrence of the human-specific Bacteroides fecal marker at an open coast marine beach: relationship to waves, tides and traditional indicators". *Environ Microbiol* 9, pp. 2038–2049. DOI: 10.1111/j.1462-2920.2007.01319.x.

[115] P.A. Bower, C.O. Scopel, E.T. Jensen, M.M. Depas, and S.L. McLellan. 2005. "Detection of genetic markers of fecal indicator bacteria in Lake Michigan and determination of their relationship to Escherichia coli densities using standard microbiological methods". *Appl Environ Microbiol* 71, pp. 8305–8313. DOI: 10.1128/AEM.71.12.8305-8313.2005.

[116] C.D. Heaney, E. Sams, A.P. Dufour, K.P. Brenner, R.A. Haugland, E. Chern, S. Wing, S. Marshall, D.C. Love, M. Serre, R. Noble, and T.J. Wade. 2012. "Fecal indicators in sand, sand contact, and risk of enteric illness among beachgoers". *Epidemiology* 23, pp. 95–106. DOI: 10.1097/EDE.0b013e31823b504c.

[117] I.U. Khan, S. Hill, E. Nowak, and T.A. Edge. 2013. "Effect of incubation temperature on the detection of thermophilic campylobacter species from freshwater beaches, nearby wastewater effluents, and bird fecal droppings". *Appl Environ Microbiol* 79, pp. 7639–7645. DOI: 10.1128/AEM.02324-13.

[118] R.L. Mohammed, A. Echeverry, C.M. Stinson, M. Green, T.D. Bonilla, A. Hartz, D.S. McCorquodale, A. Rogerson, and N. Esiobu. 2012. "Survival trends of *Staphylococcus aureus*, *Pseudomonas aeruginosa*, and *Clostridium perfringens* in a sandy South Florida beach". *Mar Pollut Bull* 64, pp. 1201–1209. DOI: 10.1016/j.marpolbul.2012.03.010.

[119] T.J. Wade, E. Sams, K.P. Brenner, R. Haugland, E. Chern, M. Beach, L. Wymer, C.C. Rankin, D. Love, Q. Li, R. Noble, and A.P. Dufour. 2010. "Rapidly measured indicators of recreational water quality and swimming-associated illness at marine beaches: a prospective cohort study". *Environ Health* 9, p. 66. DOI: 10.1186/1476-069X-9-66.

[120] C.J. Palmer, Y.L. Tsai, C. Paszko-Kolva, C. Mayer, and L.R. Sangermano. 1993. "Detection of *Legionella* species in sewage and ocean water by polymerase chain reaction, direct fluorescent-antibody, and plate culture methods". *Appl Environ Microbiol* 59, pp. 3618–3624.

[121] J. Dabrowski. 1982. "Isolation of the Shigella genus bacteria from the beach sand and water of the bay of Gdansk". *Bull Inst Marit Trop Med Gdynia* 33, pp. 49–53.

[122] E. Levin-Edens, N. Bonilla, J.S. Meschke, and M.C. Roberts. 2011. "Survival of environmental and clinical strains of methicillin-resistant *Staphylococcus aureus*

[MRSA] in marine and fresh waters". *Water Res* 45, pp. 5681–5686. DOI: 10.1016/j.watres.2011.08.037.

[123] Z.J. Mudryk, A. Kosiorek, and P. Perliński. 2013. "In vitro antibiotic resistance of Vibrio-like organisms isolated from seawater and sand of marine recreation beach in the southern Baltic Sea". *Hydrobiologia* 702.1, pp. 141–150.

[124] H. Cui, K. Yang, E. Pagaling, and T. Yan. 2013. "Spatial and temporal variation in enterococcal abundance and its relationship to the microbial community in Hawaii beach sand and water". *Appl Environ Microbiol* 79, pp. 3601–3609. DOI: 10.1128/AEM.00135-13.

[125] C. Rinke, P. Schwientek, A. Sczyrba, N.N. Ivanova, I.J. Anderson, J.F. Cheng, A. Darling, S. Malfatti, B.K. Swan, E.A. Gies, J.A. Dodsworth, B.P. Hedlund, G. Tsiamis, S.M. Sievert, W.T. Liu, J.A. Eisen, S.J. Hallam, N.C. Kyrpides, R. Stepanauskas, E.M. Rubin, P. Hugenholtz, and T. Woyke. 2013. "Insights into the phylogeny and coding potential of microbial dark matter". *Nature* 499, pp. 431–437. DOI: 10.1038/nature12352.

[126] B.P. Hedlund, J.A. Dodsworth, S.K. Murugapiran, C. Rinke, and T. Woyke. 2014. "Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter"". *Extremophiles* 18, pp. 865–875. DOI: 10.1007/s00792-014-0664-7.

[127] D. Kim, L. Song, F. Breitwieser, and S. Salzberg. 2016. "Centrifuge: rapid and sensitive classification of metagenomic sequences". *bioRxiv*, p. 054965.

[128] M.B. Jones, S.K. Highlander, E.L. Anderson, W. Li, M. Dayrit, N. Klitgord, M.M. Fabani, V. Seguritan, J. Green, D.T. Pride, S. Yooseph, W. Biggs, K.E. Nelson, and J.C. Venter. 2015. "Library preparation methodology can influence genomic and functional predictions in human microbiome research". *Proc Natl Acad Sci U S A* 112, pp. 14024–14029. DOI: 10.1073/pnas.1519288112.

[129] C.E. Shannon. 1997. "The mathematical theory of communication. 1963". *MD Comput* 14, pp. 306–317.

[130] E.H. Simpson. 1949. "Measurement of diversity". *Nature*.

[131] J.R. Bray and J.T. Curtis. 1957. "An ordination of the upland forest communities of southern Wisconsin". *Ecol Monogr* 27.4, pp. 325–349.