Taxonomic and Environmental Annotation of Bacterial 16S rRNA gene sequences *via* Shannon Entropy and Database Metadata Terms

by

Ali Zeeshan Ijaz

Hawkesbury Institute for the Environment

Western Sydney University, Australia

Thesis submitted: February 2017

A thesis submitted for fulfilment of the requirements of the degree of

Doctor of Philosophy

Dedication

I would like to dedicate my thesis to my beloved grandfather, Iqbal Ahmad and my elder sister Sidrah Ijaz, who have taught me to work hard for the things that I aspire to achieve. Thank you for your support along the way.

Acknowledgements

I am greatly indebted to my PhD supervisor, Prof. Brajesh Singh and co-supervisor Dr. Christopher Quince for identifying and streamlining the research to be undertaken for the completion of my PhD at Hawkesbury Institute for the Environment, Western Sydney University. Their professional competence, supervision and continued support in all areas of personal and professional interest have been paramount towards completion of my studies with zeal and enthusiasm.

All along I have had the opportunity to interact and enjoy the able guidance and support of other professional and competent people, with the foremost being Dr. Thomas Jeffries, to whom I extend my special thanks for his unending patience and time devotion to keep me on track and encouragement that made it possible to complete my studies. His competence and grasp of knowledge facilitated an environment that made it possible to perform my research in the right direction. I would also like to thank Jasmine Grinyer for her support in various aspects of paperwork.

Lastly, I would like to thank my family for being extremely understanding and supportive all along the study period, including my brother Dr. Umer Zeeshan Ijaz, a Research fellow at University of Glasgow. Their support and encouragement gave me the resolve to complete my PhD while being under a serious ailment.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. It contains no material previously published or written by another person. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institute.

Signed

Ali Zeeshan Ijaz

1st June 2018

Table of Contents

List of Tables

APPENDIX A

List of Figures

List of Abbreviations

%- Percent

ADONIS- A non-parametric multivariate analysis of variance

ANOSIM- Analysis of similarities

BLAST- Basic local alignment search tool

BLAT- BLAST-like alignment tool

bp- Base pair

°C- Degree Celsius

DNA- Deoxyribonucleic acid

HIE- Hawkesbury Institute for the Environment

LCA- Lowest common ancestor

MEGAN- Metagenome Analyzer

MG-RAST- Metagenomic Rapid Annotations using Subsystems Technology

NCBI- National Center for Biotechnology Information

NCBI-NR- NCBI Non-redundant database

OTU- Operational taxonomic unit

PacBio- Pacific Biosciences

pH- Potential of hydrogen

QIIME- Quantitative insights into microbial ecology

rRNA- Ribosomal ribonucleic acid

RDP- Ribosomal database project

RNA- Ribonucleic Acid

SE- Shannon entropy

Abstract

Microbial ecology seeks to describe the diversity and distribution of microorganisms in various habitats within the context of environmental variables. High throughput sequencing has greatly boosted the number and scope of projects aiming to study and analyse these organisms, with ever-increasing amounts of data being generated. Amplicon based taxonomic analysis, which determines the presence of microbial taxa in different environments on the basis of marker gene annotations, often uses percentage identity as the main metric to determine sequence similarity against databases. This data is then used to study the distribution of biodiversity as well as the response of microbial communities to stressors. However, the 16S rRNA gene displays varying degrees of sequence conservation along its length and is therefore prone to provide different results depending on the part of 16S rRNA gene used for sequencing and analysis. Furthermore, sequence alignment is primarily performed using the popular BLAST sequence alignment tool, which incurs a great computational performance penalty although newer, more efficient tools are being developed. A new approach that is fast and more accurate is critically needed to process the avalanche of data. Additionally, repositories of environmental metadata can provide contextual information to sequence annotations, potentially enhancing analysis if they can be incorporated into bioinformatics pipelines. The overarching aim of this work was to enhance the taxonomic annotation of bacterial sequences by developing a weighted scheme that utilizes inherent evolutionary conservation in the bacterial

16S rRNA gene sequences and by adding contextual, environmental information pertaining to these sequences in a systematic fashion.

In **Chapter 2**, we sought to develop a new sequence similarity metric by quantifying evolutionary conservation within the bacterial 16S rRNA gene sequences *via* Shannon entropy and comparing it against the commonly applied percentage identity. The SILVA 16S rRNA reference database (Quast et al., 2013) was used for *in-silico* comparison between both approaches by way of emulating Illumina sequencing technology using simulated datasets. The new approach showed better taxonomic annotation capability at higher taxa levels compared to the percentage identity metric, especially at family and class level. By directly utilizing the evolutionary conservation information available in bacterial 16S rRNA gene sequences, the new approach provided an effective measure of sequence similarity. This is especially important given that percentage identity metric omits this information.

In **Chapter 3**, the aim was to develop a new bioinformatics pipeline based on the Shannon entropy metric developed in Chapter 1. Analysis was performed on real amplicon datasets belonging to the sugarcane biome using the new pipeline as well as an established pipeline. Furthermore, for the new pipeline, an OTU-independent approach was followed to see if analysing each sequence could improve the overall performance of the system. Diversity results were used to compare both pipelines, under the context of OTU-based and OTU independent approaches. Results show that the new pipeline is able to effectively delineate similar ecological patterns as established pipelines. For OTU-based approaches,

TaxaSE illustrated similar alpha diversity and beta diversity patterns as QIIME at 97% OTU similarity, while for an OTU-independent approach; TaxaSE was able to provide more taxonomic annotations for the datasets.

In **Chapter 4**, the aim was to develop an environmental annotation enhancement to the bioinformatics pipeline developed in chapter 3. The SEQenv pipeline was integrated and enhanced via development of an environmentally contextual view of ecological annotation using an extension. While SEQenv only provides a list of environmental terms, the newly developed extension enabled a taxa centric approach to environmental annotations. This allowed for a contextual view of abundance of taxa on an environmental term basis as well as quantifying the distribution of various environments the taxa may come from. The results show that for sequences that are present across multiple sub-habitats, their abundance varies significantly among them. Additionally, some taxa, which did not demonstrate a cosmopolitan distribution, were found to be present in a few sub-habitats. The environmental annotation of these sequences was confirmed by previous literature available on the habitats for these microbes. Hence the new extension provided a more direct view of taxa distribution across various environments as well as illustrated environmental distribution for each taxon, significantly improving upon the SEQenv pipeline.

Overall, the new pipeline presented in this thesis provides a novel approach to annotate bacterial 16S rRNA gene sequences by way of combining a new approach to taxonomic annotation with contextual environmental information. The new

pipeline can be a valuable tool for biologists aiming to understand microbial communities in a more effective manner.

Chapter 1: General Introduction

1.1 Importance of Microbial Community Analysis

Microbes play a highly important role in sustaining life on planet Earth performing functions such as driving nutrient cycles (Venter et al., 2004) and influencing human health conditions (Handelsman, 2004). Furthermore, due to their versatility and resilience, they occupy a wide variety of environments which range from deep-sea vents, having a temperature in excess of 300°C, to rocks found far deep beneath earth's surface (Wooley, Godzik, & Friedberg, 2010) and are ubiquitous in habitats such as soil, the ocean and the mammalian gut. Determining the diversity and taxonomic composition of microbial communities is a central task in every project that aims to understand the impact of microbial communities on environmental systems and the factors, which control microbial diversity.

Taxonomic profiling of microbial communities is used to examine the species composition and relative abundance of the various bacteria that are present for a given habitat. Despite the widespread distribution and ecological importance of microbes, very little is known about their biology, given that only a small fraction can be cultured under laboratory conditions (Nikolaki & Tsiamis, 2013) where standard culturing techniques account only for 1% or less of bacterial diversity in most environmental samples (Riesenfeld, Schloss, & Handelsman, 2004). However, next generation DNA sequencing technology has greatly boosted the number and scope of ecological projects and produces a huge amount of microbial

data without the need of culturing. However, it produces short read sequencing data of a few hundred base pairs in length. This illustrates the need for enhanced and fast approaches towards analysis but also appropriate taxonomical identification based on more discriminatory approach than current methods, especially as more and more data is being generated.

1.2 Taxonomic Annotation using Conserved Marker Genes

Measuring species diversity is often the first step towards understanding the microbial community present in an environmental sample. Taxonomic annotation helps categorize and quantify microbial diversity in terms of species richness and relative abundance (Knights et al., 2011), using diversity indexes such as Shannon diversity and Simpson diversity (Simpson, 1949; Whittaker, 1972). Researchers have relied on the 16S rRNA gene, a 1500 to 1600 bp long sequence, the usage of which was pioneered by C. R. Woese, due to its presence across all prokaryotic species, including bacteria. It displays enough sequence diversity for phylogenetic classification and assessing the genetic diversity of environmental samples (Woese, 1987). The gene has been widely used for sequencing and identification of many bacterial isolates and to profile uncultured microbial communities from diverse habitats (Gillian C. Baker, Gaffar, Cowan, & Suharto, 2001; Grosskopf, Janssen, & Liesack, 1998; McInerney, Wilkinson, Patching, Embley, & Powell, 1995).

Given the immensely important role microbes play in ecosystem function and biochemical cycles (Falkowski, Fenchel, & Delong, 2008), various surveys of diversity have been conducted using 16S rRNA gene sequences to elucidate the impact of microbial communities on their environment and habitats. This includes exploration of the diversity and functional characteristics of soil microbial communities across various biomes (Noah Fierer et al., 2012) and the geographical distribution of marine bacterial communities (Ghiglione et al., 2012). Furthermore, changes in environmental factors such as temperature have been shown to produce variation in the structure of bacterial communities in soil (Xiong et al., 2014) and especially permafrost (Mackelprang, Saleska, Jacobsen, Jansson, & Taş, 2016), where an increase in temperature due to climate change may likely result in significant losses in soil carbon (McCalley et al., 2014) and therefore reshape the environment (Gibbons & Gilbert, 2015).

Microbes also play a significantly important role in human health. Various studies have been conducted to determine the composition of gut microbiota and the impact they have on various aspect of human health. They have been shown to be crucial for protection against food allergies (Stefka et al., 2014) as well as autoimmune disorders (Hooper, Littman, & Macpherson, 2012). Additionally, recent research has revealed the role gut microbiome play in development of the central nervous system (Sharon, Sampson, Geschwind, & Mazmanian, 2016). Furthermore, various diseases such as asthma (von Mutius, 2016), and rheumatic autoimmune diseases (Coit & Sawalha, 2016) are directly impacted by the composition of the human microbiome, in addition to being an important

contributing factor to the development of gastric cancer (Wroblewski, Peek, & Coburn, 2016).

The tree of life where the phylogenetic relationships between bacteria, archaea and eukaryotes, are reconstructed based on the small ribosomal 16S rRNA is shown in Figure 1-1 and illustrates the evolutionary relationship between these different kingdoms. Culture free 16S rRNA gene sequence-based tools have significantly expanded our view of microbial diversity, where polymerase chain reaction or PCR amplification of 16S rRNA gene sequences has provided great insight into our understanding of microbial communities, as it enabled sequencing of those microbial genes, which are as yet uncultivable or exist in extreme habitats (G. C. Baker, Smith, & Cowan, 2003).

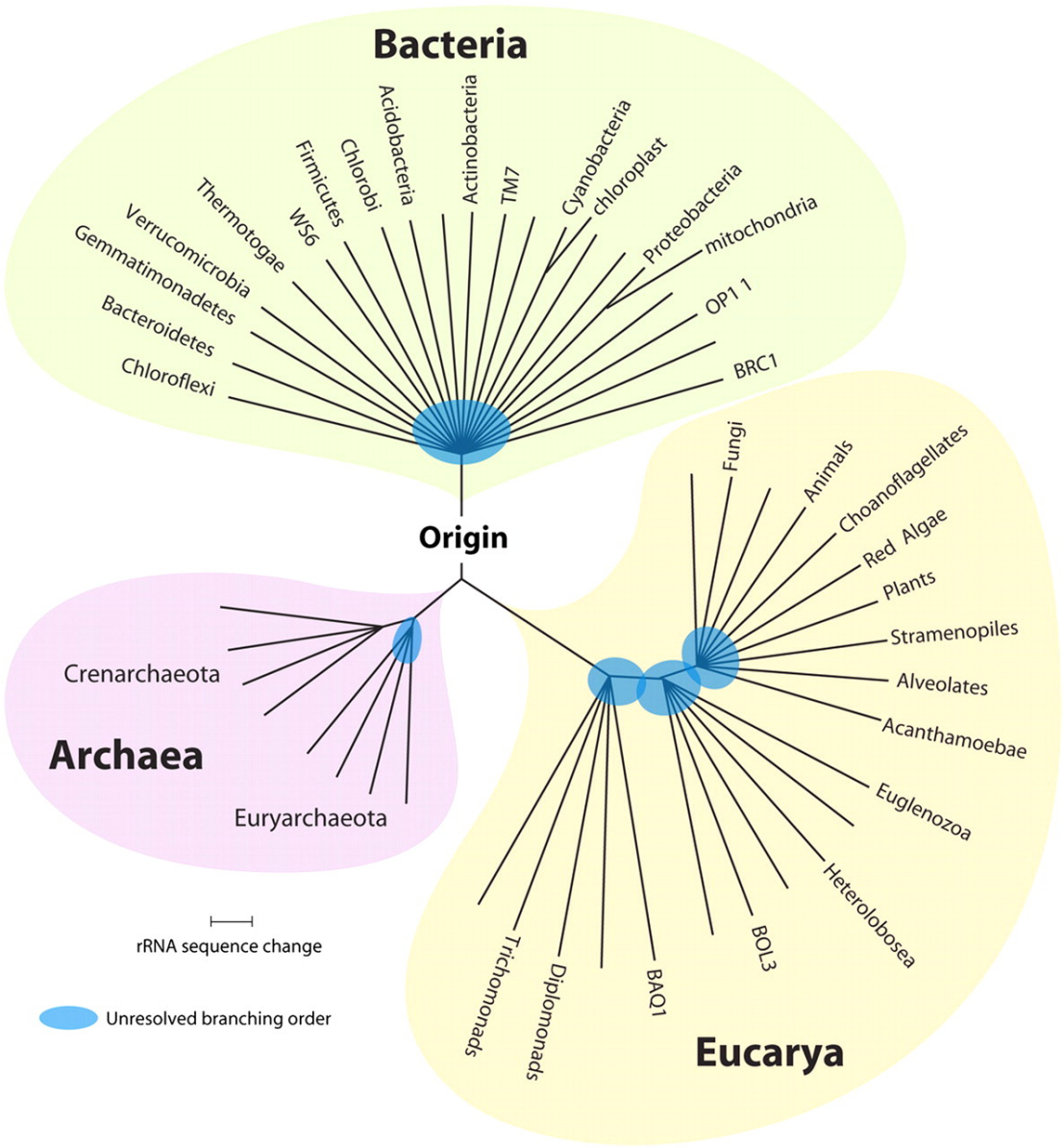


Figure 1–1: Phylogenetic Tree of Life, illustrating the three main branches of Bacteria, Archaea and Eucarya. Reproduced from (Nair, 2012)

1.2.1 Amplicon Sequencing of 16S rRNA Gene Sequences

Ideally, the whole 16S rRNA gene should be sequenced for the study of sequences isolated directly from the environment, also known as ecogenomic surveys, but currently the read length of next generation sequencing technologies precludes

this approach and thus most surveys aim at characterisation of selected hyper-variable regions as these can effectively distinguish between different taxa (Santamaria et al., 2012), due to the fact that nucleotides in these regions change more rapidly between sequences. Nine hypervariable regions (V1-V9) are present (Figure 1-2), each exhibiting a different degree of sequence diversity and no single region may differentiate among all bacteria (Chakravorty, Helb, Burday, Connell, & Alland, 2007). Hence to analyse the taxonomic content of an environmental sample, biologists have typically used amplicon sequencing in which a particular variable region is amplified, as conserved regions in most bacterial DNA sequences flank these regions, which enables PCR amplification of target sequencing using primers. DNA fragments of the 16S rRNA gene can be selectively amplified from mixed DNA, leading to significant improvement in sequencing throughput (Amann, Ludwig, & Schleifer, 1995), as only genes of interest are amplified for DNA sequencing. There is already a number of primers being used for amplification and sequencing, with some of them being referred as universal primers as they provide coverage of a majority of 16S rRNA gene sequences (Watanabe, Kodama, & Harayama, 2001).

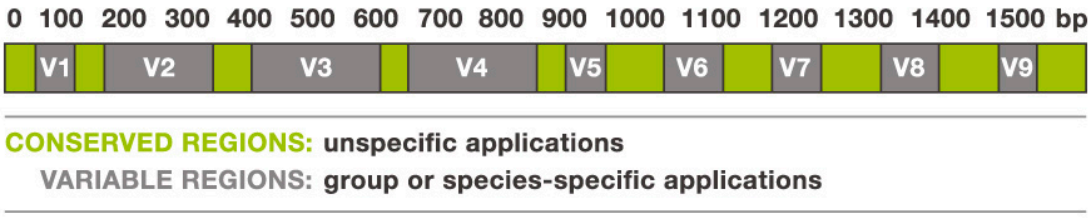


Figure 1–2: 16S rRNA gene composition where variable regions are highlighted in grey while conserved regions of the gene are illustrated in green. Reproduced from (Alimetrics)

A second-generation sequencing system, the Illumina platform, is currently preferred for amplicon sequencing due to its low cost and deeper coverage with small read lengths, going up to 250bp, and consequently is the most widely used sequencing platform (Logares et al., 2014). Given that the read length is increasing (Laver et al., 2015), addition of other regions for analysis may become more useful as more information can now be derived. Other second-generation sequencing platforms include Ion Torrent, which generates a read length of 200 bp and more (Quail et al., 2012) but suffers from homopolymer errors and Roche 454, which generates a read length of 800 bp or more (Loman et al., 2012) but is expensive to run.

Newer sequencing platforms have been developed, known as third generation sequencing systems. This include PacBio that can produce an average read length of 2500bp or more, and some longer reads reaching 10000bp (Au, Underwood, Lee, & Wong, 2012), as well as Oxford Nanopore, where read length can be tens of

kilobases on average (Laver et al., 2015). However, these platforms exhibit a high error rate, which can reach up to 40%, and may require data from second-generation sequencing system, such as Illumina to bring the error down to a respectable level, which then reduces their throughput. Hence, the Illumina platform is still applicable for sequencing due to its high throughput, very low error rate (2%) and for use in assisting third-generation sequencing technologies (Au et al., 2012).

1.2.2 Reference Databases

Various new taxonomic groups have been discovered as a result of ecogenomic surveys (Hugenholtz, Pitulle, Hershberger, & Pace, 1998; McInerney et al., 1995; Nielsen et al., 1999) and consequently the list of 16S rRNA gene sequences has been growing, with many of 16S rRNA reference sequences in publically available databases, such as SILVA (Quast et al., 2013), Ribosomal Database Project (Cole et al., 2014) or Greengenes (DeSantis et al., 2006). Reference sequences are DNA sequences considered as representative of the species they belong to and typically follow a stringent process to determine their validity. These databases consist of millions of sequences and are being readily used in many ecological projects to determine the taxonomic diversity of given environmental samples such as human intestinal ecosystem (Ritari, Salojärvi, Lahti, & de Vos, 2015), deep sea habitat (Sogin et al., 2006) and soil (Johannes Rousk et al., 2010). Both aligned, where reference sequences are aligned together, and non-aligned versions, where

reference sequences are provided without alignment, of these databases are available from them.

SILVA provides quality checked and regularly updated databases of both small (16S, 18S) and large (23S, 28S) ribosomal RNA gene sequences belonging to bacteria, archaea and eukaryotes. Two subsets are available; the first named as SSU-Parc, which is intended for biodiversity analysis and SSU-Ref, which consists of nearly full-length sequences, is intended for designing probes and phylogenetic analysis (Quast et al., 2013; Santamaria et al., 2012). The database is built upon the ARB software, which is used for sequence database management and analysis (Ludwig et al., 2004).

Greengenes provides phylogenetic classification of 16S rRNA gene sequences from GenBank (Benson et al., 2012) and uses taxonomy from NCBI and Ribosomal Database Project (Cole et al., 2014) as well as information provided by independent curators such as Phil Hugenholtz (Hugenholtz, 2002), Wolfgang Ludwig (Amann et al., 1995), and Norman Pace (Pace, 1997). The database has high quality sequences and is compatible with ARB software.

The Ribosomal Database Project or RDP database provides another source of taxonomically annotated reference 16S rRNA gene sequences, which are also available from International Nucleotide Sequence Database Collaboration (INSDC) (Balvočiūtė & Huson, 2017). The database consists of both bacterial and archaeal sequences with most of the sequences being incomplete, which are derived from sequencing PCR amplification products (Cole et al., 2014). Other

databases include NCBI and GenBank (Benson et al., 2012). NCBI provides a non-curated, authoritative classification of sequences and taxonomy which includes prokaryotic and eukaryotic species (Santamaria et al., 2012). The choice of the database is dependent on the tools provided as well as the quality and quantity of sequences. Considering that the taxonomy may differ between these databases, the selection of database plays an extremely important role in downstream analysis. Greengenes and SILVA databases are most prevalent for 16S rRNA gene sequence based analysis.

1.2.3 Preprocessing of DNA Sequence Data

Before accurate analysis can take place, the sequences need to undergo preprocessing to remove spurious data. This includes chimera detection and removal, sequence trimming and merging. Chimeras are artificial recombinants between two or more sequences and are formed during PCR amplification, where prematurely terminated DNA fragments re-anneal to another DNA. Given that the presence of these sequences makes it difficult to differentiate between real sequences from recombinants as the breakpoints can occur at any location and the next generation sequencing platform generate short sequences, making it harder to distinguish the chimera from its parents. Hence an overestimation of microbial diversity is observed (M. Kim et al., 2013). Hence their removal is important for proper analysis. Popular tools include UCHIME (Edgar, Haas, Clemente, Quince, & Knight, 2011), and ChimeraSlayer (Haas et al., 2011). Removal of chimera is hence essential for downstream analysis.

Raw DNA sequences obtained from sequencing machines vary in per-base quality, with base-call errors being observed in these sequences. The quality of a sequence is determined using a PHRED score, which represent the probability of a base call error at every nucleotide base in the sequence and is observed to decrease across the length of the sequence read (Shrestha et al., 2014). Various tools, such as Seqtk (Li, 2017), QTrim (Shrestha et al., 2014) and Fastx (Blankenberg et al., 2010) are used for trimming low quality bases from the sequences and are invaluable in reducing erroneous data.

Next generation technologies like Illumina sequencing produce paired-end reads which can be merged together to form a longer sequence. The overlapping region between these reads can correct for sequencing errors and improve the quality of the resulting DNA sequence (J. Zhang, Kobert, Flouri, & Stamatakis, 2014). FLASH (Magoc & Salzberg, 2011) and PANDAseq (Masella, Bartram, Truszkowski, Brown, & Neufeld, 2012) are some of the tools used for paired-end read merging. Furthermore, downstream analysis also benefit from this approach as accuracy improves with length of the sequence.

Lastly, newer noise reduction methods can further reduce spurious reads and improve downstream analysis. Some of these tools, such as DADA2 (Callahan et al., 2016) can model and correct Illumina-sequenced amplicon errors. It can resolve differences of as little as a single nucleotide and produces fewer incorrect sequences than other methods.

1.2.4 Sequence Aligners

Sequence alignment is an integral part of a reference-based taxonomic annotation pipeline where alignment is performed between the reference database and query sequence. As reads are aligned against reference sequences, a clearer picture of microbial species composition and abundance is realised (Reinert, Langmead, Weese, & Evers, 2015).

One of the earliest aligners used was the BLAST algorithm (Altschul, Gish, Miller, Myers, & Lipman, 1990). Useful for both nucleotide and protein sequences, BLAST has been used in a wide variety of ecological projects. However, the advent of next-generation sequencing has led to an avalanche of sequence data (Kostadinov, 2011), which needs to be analysed in a fast, effective manner. BLAST was found to be not efficient enough to analyse these datasets and therefore new sequence aligners continue to be developed (Reinert et al., 2015). These aligners are many times faster than BLAST and are being readily used in various bioinformatics pipelines. Notable examples include Usearch (Edgar, 2010), bowtie2 (Langmead & Salzberg, 2012) and BLAT (Kent, 2002). These aligners use novel approaches to sequence alignment, by way of k-mer, which are small substrings of length k and are computationally less expensive for sequence analysis, and indexing based approaches that are computationally advantageous (Mielczarek & Szyda, 2016).

1.2.5 The Percentage Identity Metric

Given the large number of sequences being produced and the computational requirements necessary to analyse them, a preliminary step is performed where DNA sequences are aligned and compared and those sequences that are similar to each other are clustered together, and classified as belonging to the same operational taxonomic unit or OTU. Most software that perform taxonomic annotation align these 16S rRNA OTUs against database reference sequences in which the sample sequence is compared with taxonomically annotated reference sequences (Santamaria et al., 2012). Sequence similarity is determined by a percentage identity metric where sequences are scored on the number of matches between reference and query sequences and penalised for any gaps in the alignment (Edgar, 2010). A match is scored where the specific nucleotide base on both query sequence and reference sequence match, and a mismatch is where these are different. Insertions and deletions are accounted by gap in the reference sequence and in the query sequence respectively. In this context, typically 99% sequence similarity is considered as a threshold for species while 97% similarity is for genus level, with family at 95%, order at 90%, class at 85% and finally phylum at 80% (Drancourt et al., 2000; Lanzen et al., 2012).

1.2.6 Popular Taxonomic Annotation Pipelines

Popular tools that perform taxonomic annotation of microbial communities include MG-RAST (Aziz et al., 2008), MEGAN (Huson, Richter, Mitra, Auch, & Schuster, 2009), QIIME (Caporaso et al., 2010) and MOTHUR (Schloss et al., 2009). Almost all of these have traditionally been dependent on the BLAST algorithm

(Altschul et al., 1990) for sequence alignment, although newer versions have begun a shift to other algorithms such as USEARCH (Edgar, 2010) and BLAT (Kent, 2002), e.g. MEGAN can use DIAMOND (Buchfink, Xie, & Huson, 2015), which is significantly faster, due to the significantly higher computational requirement for BLAST and a need for high throughput.

MG-RAST provides an online service for phylogenetic and functional annotation of metagenomes. For rRNA sequences, the service uses the QIIME pipeline (Caporaso et al., 2010). MEGAN, which stands for Metagenome analyser, is a stand-alone tool that is primarily aimed towards taxonomic annotation of metagenomes and uses the NCBI-NR database and BLAST, in conjunction with a Lowest Common Ancestor (LCA) algorithm, for taxonomic assignments (Huson et al., 2009). QIIME or Quantitative Insights Into Microbial Ecology is an extensive suite of bioinformatics tools for analysis of microbial communities, which combines several taxonomic assignment tools like UPARSE and the RDP annotation tool (Caporaso et al., 2010; Cole et al., 2014; Edgar, 2013). For 16S rRNA gene sequences, the Greengenes database (DeSantis et al., 2006) is typically used for taxonomic annotation, although other databases including SILVA (Quast et al., 2013) are also available for QIIME. A wide variety of statistical analysis can be performed in the software. Lastly, similar to QIIME, MOTHUR (Schloss et al., 2009) is another suite of bioinformatics tools for the annotation of taxonomic marker genes and includes a variety of analysis tools. These tools primarily use the percentage identity metric for sequence similarity measurements. Online services may be slow depending on the usage. Lastly, environmental annotation capability is not available under these pipelines.

1.2.7 Phylogenetic Placement Algorithms

Algorithms such as pplacer (Matsen, Kodner, & Armbrust, 2010) or RAxML EPA (Berger, Krompass, & Stamatakis, 2011) provide means to explore evolutionary origin of query sequences by way of phylogenetic placement. By utilizing models of rate heterogeneity among sites such as Gamma-distributed rate heterogeneity (Yang, 1994) or the CAT model of site-specific character frequencies (Lartillot & Philippe, 2004), these tools are able to determine the evolutionary relations of the query sequences with other sequences in a phylogenetic tree and therefore perform fine-scale analysis of sequences for comparative and evolutionary information using a phylogenetic distance metric (Matsen et al., 2010). The CAT model is faster and uses less memory than Gamma-distributed model, while producing slightly better Gamma likelihood values, making it computationally feasible to analyse large trees (Stamatakis, 2006). Quick and efficient, these tools are able to place thousands of query sequences onto a phylogenetic tree in linear time and memory complexity and hence are suitable for analysis of large-scale metagenomic datasets.

However, these methods differ from traditional taxonomic annotation approaches, as they do not assign names to query sequences. Furthermore, phylogenetic placement is designed to work with a single reference phylogenetic tree, built using a single alignment and hence is only suitable for single gene based analysis.

1.3 Environmental Annotation of Sequences

Microorganisms are genetically diverse and occupy every known habitat. While taxonomic annotation tools answer the question "who is there?" they are unable to provide environmental context for these taxonomic annotations. This information is especially important given that microorganisms are found in a variety of environments (generalists), while others occupy a specific niche defined by key environmental parameters (specialists) (Kuenen, 1983; Monard, Gantner, Bertilsson, Hallin, & Stenlid, 2016). Understanding how the environment selects particular taxa and the diversity patterns that emerge as a result of environmental filtering, can dramatically improve our ability to analyse any environment in depth. Furthermore, this will improve our knowledge on how the response of different taxa can impact each other and ecosystem functions, especially in the context of Baas-Becking hypothesis, which states that everything is everywhere but the environment selects (Baas-Becking, 1934; De Wit & Bouvier, 2006). Members of rare taxa account for most of the observed phylogenetic diversity (Sogin et al., 2006), and become more abundant if the environmental conditions favour their growth (Shade et al., 2014). This is because the organisms can experience conditions that are not optimal for their growth and therefore enter in a state of reversible dormancy (Lennon & Jones, 2011). This leads to microbes exhibiting biogeographic distribution patterns which differ from patterns observed for plants and animals (Xia et al., 2016).

Most of the work investigating microbial biogeography has been site-specific and logical environmental factors, rather than geographical location, may be more

influential on microbial diversity (Fierer & Jackson, 2006). Furthermore, the level of nutrients also determines the growth and diversity of organisms present in a habitat, such as where certain taxa are observed on the basis of whether the microorganism is oligotrophic or copiotroph in nature (Koch, 2001). Copiotrophs have high growth rates when nutritional conditions are abundant while oligotroph demonstrate slower growth rate and may even outcompete copiotrophs when the level of nutrients available is low, based on the flexibility available in their genomes (Fierer, Bradford, & Jackson, 2007).

1.3.1 Factors Influencing Microbial Community Composition

Vellend (2010) proposed that mechanisms, which shape the composition and diversity of microbial communities could be divided into four classes termed speciation, selection, dispersal and ecological drift. Speciation adds more species diversity over time, selection modifies the relative abundance of taxa on the basis of the survival and reproducibility capability of these species, dispersal of established species to a new location brings change in the community composition depending on the local conditions and finally ecological drift where chance demographic fluctuations can lead to a change in species abundance (Hanson, Fuhrman, Horner-Devine, & Martiny, 2012; Vellend, 2010).

While various environmental factors determine the bacterial diversity in a biome, key factors tend to be better predictors of microbial diversity. For example, soil pH level was the best predictor of diversity and richness in various soil samples

collected from North and South America (Fierer & Jackson, 2006). Furthermore, environmental conditions such as climatic or land cover characteristics influence bacterial community structure as well at regional and global scales (Xia et al., 2016). Environmental parameters drive the composition and structure of microbial communities in all habitats (Jeffries et al., 2011), with parameters such as soil nutrient availability (Broughton & Gross, 2000), salinity and pH (Lozupone & Knight, 2007; Alban Ramette & James M. Tiedje, 2007) and lastly plant diversity (Stephan, Meyer, & Schmid, 2000) have also been found to influence the microbial community composition and diversity in soil ecosystems, suggesting microbial communities respond to multiple environmental factors. These factors together define a niche occupied by particular species or community. However, the science behind niche formation and community assembly remains poorly understood.

A full understanding of the role of environmental drivers of microbial diversity can only be realised if associated metadata related to geographical or environmental information can be exploited (Alban Ramette & James M. Tiedje, 2007). Knowing just the taxonomy of the species present may not be enough, given the need to understand the niche and controlling variables of microorganisms. A complex combination of historical factors such as dispersal limitation and past environmental conditions significantly influence present-day groupings of microbes in addition to overall contemporary habitat characteristics (Dinsdale et al., 2008) as well as changes in environmental parameters. Understanding these parameters is critical to understand evolution, community assembly and microbial ecology.

1.4 Knowledge Gaps

Given that gene sequences evolve under evolutionary constraints, certain portions of these sequences tend to be more variable than other, conserved regions (Woese, 1987), henceforth considered as evolutionary conservation. While popular bioinformatics pipelines are being used for many ecological projects, they mainly use the percentage identity metric for determining sequence similarity. Metric selection has a significant impact on the analysis being conducted, as all downstream analyses, including diversity measurements, depend on sequence alignments and the database used. While hypervariable regions are primarily used to differentiate between different bacterial species, substantial difference is also present in non-hypervariable regions of the 16S rRNA gene (Stackebrandt & Goebel, 1994), which can also play a role in differentiating between various bacteria. Furthermore, closely related species that differ by a few nucleotide bases can be erroneously considered identical and may require comparing specific locations on the 16S rRNA gene sequences in order to determine their differences thereby significantly reducing taxonomic resolution at species or sub-species level (Fox, Wisotzkey, & Jurtshuk, 1992). Furthermore, the selection of the primer for amplicon sequencing and the targeted variable regions also directly impact the analysis of microbial communities, with different primers producing different abundances of taxa (Fredriksson, Hermansson, & Wilen, 2013). This represents the limitation of 16S rRNA gene sequence analysis and therefore there is a great need to improve the analytical capabilities of any bioinformatics pipeline that aims to perform taxonomic annotation, especially at the lower taxonomic levels, with respect to accuracy and throughput of the analysis.

With the drive to enhance the analysis of microbial community data, contextual information related to the environment they exist in, is lacking in popular pipelines. This information can be transformative in understanding the microbes and the role they play in the environment in a more thorough fashion. How environmental parameters drive diversity patterns and how diversity is partitioned by habitat, provide a contextual view that would otherwise not be observed when only using taxonomic annotation. Thus, the following knowledge gaps needs to be addressed:

- The percentage identity metric does not account for variability at any match or mismatch location and therefore does not fully exploit the evolutionary conservation and variability inherent in the gene (Woese, 1987) as the degree of variability changes across these hypervariable regions (Figure 1-3). The Shannon index here denotes Shannon entropy values, which quantifies the variability across all nucleotides in the 16S rRNA gene sequences and were calculated using frequencies of the four nucleotides and gaps (Seedorf, Kittelmann, Henderson, & Janssen, 2014). The evolutionary distance between sequences as determined by percentage identity is therefore an underestimation and results in less accurate similarity scoring (Woese, 1987).

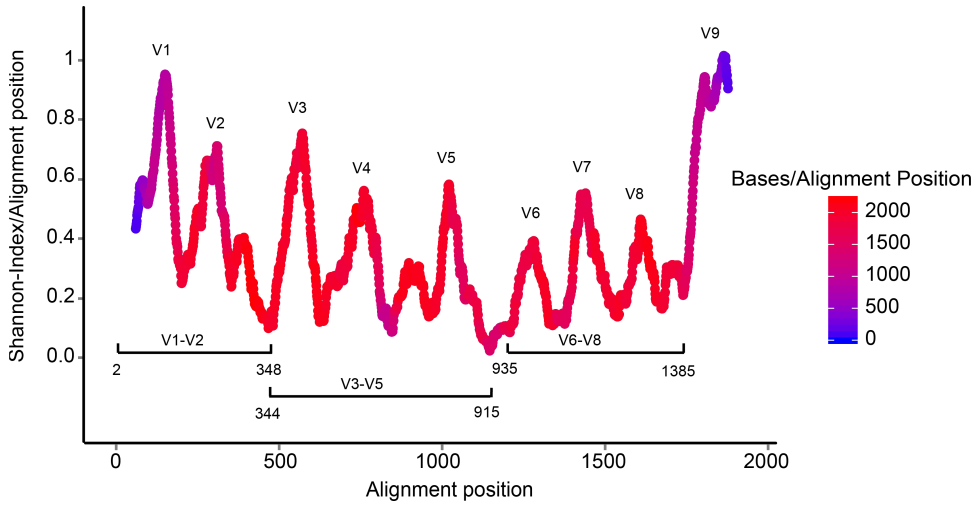


Figure 1–3: Per base variability present across the whole 16S rRNA gene quantified using Shannon-Index. Reproduced from (Seedorf et al., 2014)

- While geographical data can be used for microbial sequences, they may not be representative, as similar microbes are more likely to found in the same environments across different geographical locations (Fierer & Jackson, 2006; Xia et al., 2016). As microbial communities respond to multiple environmental factors, new data obtained from next-generation sequencing along with contextual metadata provide an important opportunity to address the need to identify the origin of sequences from a particular environment or niche. Hence an effective approach to environmental annotation of sequences is needed.

1.5 Aims and Objectives

To address the above knowledge gaps, the overarching aim of the thesis was to develop an enhanced bioinformatics pipeline for taxonomic and environmental annotation of 16S rRNA bacterial sequences, as a lot more reference sequences are available for bacteria than archaea. For taxonomic annotation, a novel metric was developed to replace percentage identity, while environmental annotation of the sequences was achieved using the SEQenv pipeline (Sinclair et al., 2016) and then extending it to provide taxa abundance within different environmental terms.

Chapter 2 aimed to develop a new sequence similarity measure that quantify and utilizes the inherent evolutionary conservation within the 16S rRNA gene sequence, which has not been utilized so far, in order to enhance taxonomic annotation. This chapter examined the hypothesis that the new similarity measure metric based on Shannon entropy would provide more robust data in comparison to percentage identity for taxonomic annotation of sequences. The new metric demonstrated improved annotation capability at higher taxa levels. The objectives of this chapter were:

a) To develop a new sequence similarity measure, which utilized evolutionary conservation within 16S rRNA gene sequences directly.
b) To determine if the new metric can be used instead of the industry standard percentage identity measure and if there are advantages to the new approach.

Chapter 3 aimed to develop a taxonomic annotation pipeline, which could use the newly developed sequence similarity metric described in chapter 2 and to test the pipeline on real amplicon datasets and comparison to QIIME. The new pipeline produced comparable ecologically important patterns. Furthermore, following an OTU independent approach provided more taxonomic annotations for sequences. The objectives of this chapter were:

a) To develop a taxonomic pipeline on the basis of the novel metric for sequence similarity developed in chapter 2.
b) To assess the applicability of the new pipeline on the analysis of real amplicon datasets belonging to samples from a sugarcane environment.
c) To assess if similar ecological patterns in comparison to QIIME are generated with the new pipeline.

Chapter 4 aimed to enhance the taxonomic annotation system developed in chapter 3 with the integration of an extended SEQenv pipeline (Sinclair et al., 2016) as a means to provide environmental annotation of 16S rRNA sequences. The extension to SEQenv was developed to provide a more contextual view into the environmental annotation of these sequences. This would be relevant to biologists determining taxa distributions for particular environments as SEQenv itself generates a word cloud only at the dataset level, listing various environmental terms acquired from the analysis and hence is limited to the dataset level. This chapter tested the hypothesis that environmental annotation could enhance analysis of microbial communities and the annotations generated were in accordance with prior knowledge in the literature about the habitats the microbes belong to. The objectives of this chapter were:

a) To integrate SEQenv into the taxonomic annotation system developed in chapter 3.
b) To assess the enhanced SEQenv system on real, amplicon datasets to validate the software and determine if the environmental annotations were in accordance with literature.

The outcome of the thesis included a single bioinformatics pipeline that produced enhanced annotation of bacterial sequences by combining high-resolution taxonomic annotation with contextual environmental annotation, which differentiated it from other pipelines. This would serve as a significantly important tool for any biologist aiming to understand microbial communities in a more effective manner.

Chapter 2: Exploiting The Evolutionary Conservation of 16S rRNA Gene via Shannon Entropy

2.1 Introduction

Microbes underpin key ecosystem services such as primary production, climate regulation (Handelsman, 2004) and elemental cycles (Venter et al., 2004), and are capable of living in diverse environments (Wooley et al., 2010). However very little is known about their biology as only a small fraction can be cultured under laboratory conditions (Nikolaki & Tsiamis, 2013) whereby standard culturing techniques account for 1% or less of bacterial diversity in most environmental samples (Riesenfeld et al., 2004).

Ecogenomics study, which seeks to understand the diversity and interactions of microbes in their natural habitats (Chapman, Robalino, & F. Trent III, 2006), is a rapidly growing field of research that aims at studying uncultured organisms via their nucleic acid sequences to understand the true diversity of microbes, their function and distribution in a variety of environments (Huson et al., 2009). Many environments have been the focus of ecogenomics studies, including soil, the oral cavity, feces, and aquatic habitats (Riesenfeld et al., 2004). The field has been driven by the advent of high throughput sequencing where genomic information is acquired directly from the microbial communities in their natural environment, with a drastic reduction in the cost of sequencing (Morgan & Huttenhower, 2014).

Culture independent studies have been used to characterise microbial communities where next-generation sequencing has been used to sequence DNA from samples from complex environments and habitats. This requires preprocessing steps such as removal of noise from sequencing data, which is due to wrong base calls, substitution errors as well as insertion and deletion of single bases (Dohm, Lottaz, Borodina, & Himmelbauer, 2008). Furthermore, error rates tend to increase along the read length for Illumina sequencing platform (Cox, Peterson, & Biggs, 2010), while newer technologies such as Oxford nanopore and PacBio exhibit error patterns of context-specific mismatches and homopolymer indels, with a high error rate that can reach 40% in some cases (Weirather et al., 2017). This is then followed by multiple sequence alignment for generation of Operational taxonomic unit, at needed similarity such as 97% (Drancourt et al., 2000). For an environmental sample, the number of annotated OTUs and relative abundances are considered as being representative of actual diversity, however identification of total diversity requires selection of an appropriate sample size. For this purpose, estimate of species richness such as rarefaction curves are used (Barriuso, Valverde, & Mellado, 2011).

Sequencing of 16S rRNA amplicons primarily uses short reads, representing a specific region of a gene while shotgun sequencing may use whole length of the genome in small fragments, which can be analysed individually or used to construct overlapping contigs. Analysis requires a significant amount of time, typically a day or more for taxonomic annotation depending on computational resources.

The underlying scoring scheme behind sequence similarity is currently percentage identity, a simple distance based approach which doesn't fully utilize the inherent variation in evolutionary conservation within 16S rRNA gene sequences, as every base is considered equal with respect to matches and mismatches and positions of these matches and mismatches are not essential (Fox et al., 1992; Stackebrandt & Goebel, 1994). This is important in the context that certain regions of the 16S rRNA gene sequences are considerably variable while others are relatively conserved, and the degree of variability is not constant (Chakravorty et al., 2007; Stackebrandt & Goebel, 1994), due to the fact that the 16S rRNA gene undergoes evolutionary changes depending on various constraints, which leads to some portions of the sequence to be highly variable while other portions to be conserved, as these conserved regions are important for the function of the 16S rRNA. (Hence known as evolutionary conservation). C. R. Woese in his work *"bacterial evolution",* stated that the distance based approach underestimates the true evolutionary distances between sequences as different nucleotide positions on sequences are changing at different rates (Woese, 1987). These represent the limitations of 16S rRNA gene sequence analysis primarily due to the selection of percentage identity as the determinant of sequence similarity.

The valuable information contained within the 16S rRNA gene sequence itself can be utilized for better understanding of sequence similarity (Stackebrandt & Goebel, 1994) and to achieve a more effective similarity measure than the current percentage identity, as discriminatory information is present in the 16S rRNA gene sequences that can be used to distinguish between various sequences. The evolutionary conservation within 16S rRNA can be determine *via* Shannon

Entropy, which quantifies the uncertainty in a random variable and hence is appropriate to determine the variability in a nucleotide position. It allows quantitative assessment of variability across the whole of 16S rRNA gene sequence in the context of an aligned bacteria database and can in turn be used to develop a similarity metric that can be more applicable to taxonomic annotation. In fact, C. R. Woese (1987) stated that determining the pattern of change at given positions in 16S rRNA gene sequence may optimise analysis (Woese, 1987). It has been utilized in other tools which utilize taxonomic marker genes such as oligotyping, which looks at nucleotide base variation within an individual OTU (Eren et al., 2013) by relying on entropy information generated through the analysis of sequences that were initially mapped onto the same taxon. Minimum Entropy Decomposition or MED, extends oligotyping via development of an unsupervised algorithm, which partitions large datasets into ecologically and phylogenetically useful units (Eren et al., 2015). In contrast to oligotyping, MED does not require any clustering of sequences into OTUs or user supervision and can be applied to whole datasets instead of only closely related sequences. Given the potential advantage of elucidating diversity, oligotyping has been used across various studies, including microbial biogeography (Cloutier, Alm, & McLellan, 2015; V. T. Schmidt et al., 2014) and microbe disease linkage (Eren et al., 2011). However, these tools only select few nucleotide positions in variable regions of the 16S marker gene and hence may not be fully utilizing all of the available entropy information present.

This study aims to address these issues by developing a novel approach to measure sequence similarity by directly using evolutionary conservation

information *via* Shannon entropy. This can then be used to enhance taxonomic annotation as sequence similarity plays an important role in reference based taxonomic annotation, where query sequences are compared against reference sequence hits on the basis of similarity (Wooley et al., 2010). Given that most of taxonomic annotation pipelines such as QIIME (Caporaso et al., 2010), MG-RAST (Aziz et al., 2008) and MEGAN (Huson, Auch, Qi, & Schuster, 2007) are dependent on percentage identity for sequence similarity measurements, an improvement in this context would result in better downstream analysis. Therefore, this chapter examines the hypothesis that a new metric based on Shannon entropy based similarity measure would provide more robust data in comparison to percentage identity for taxonomic annotation of sequences.

Furthermore, newer sequencing technologies such as PacBio (Mosher et al., 2014) and Oxford Nanopore (Laver et al., 2015) can now enable full length sequencing of 16S rRNA gene. This would in turn make the new approach more beneficial as more of the evolutionary conservation information can be used for sequence similarity measure then what is possible at the moment.

2.2 Materials and Methods

2.2.1 Shannon Entropy

To build upon existing algorithms for taxonomic identification we utilized Shannon Entropy. Every location in a DNA string can be taken as a random variable having the aforementioned nucleotide values. Entropy is a measure of the uncertainty in a random variable. In this context, the term usually refers to Shannon entropy, which quantifies the expected value of the information contained in a message (Shannon, 2001). For DNA sequences, every base location can be considered as a random variable.

In order to calculate Shannon entropy for a number of sequences, the following formula is used:

$$H = -\sum p(x) \log p(x)$$

The base of the Log function is typically 2, e or 10, though any positive real number not equal to 1 can be used. p(x) denotes probability of a variable **x**. In this context, **x** can be A, T, C, G, N and other nucleotides as well as gaps. In the context of DNA sequences, a variable region may have multiple nucleotides, each with low probability of occurrence while for conserved regions; a single nucleotide (i.e. adenine, or A) may have the highest probability with the remaining nucleotides (i.e. C, G, T) having low probability. Hence, evolutionary conservation and

variability within the 16S rRNA gene sequence is quantitatively assessed using this approach.

2.2.2 Generation of Vector Data

An aligned database of 16S rRNA sequences was used to quantitatively assess and calculate entropy across the whole 16S rRNA sequence. It is a common practice to ignore hyper-variable regions when generating a deep-level phylogeny. However, when assigning sequences to OTUs or using phylogenies for community-based hypothesis tests, the fine level of detail contained within these variable regions is significant and should not be removed (Schloss, 2009). Of the available aligned 16S rRNA gene reference databases, SILVA Release 123 aligned database (Quast et al., 2013) was selected as it contained alignment information for the whole of 16S rRNA reference sequences. For the purpose of this study, only the bacterial 16S rRNA sequences were used.

The database was taken as a matrix **M** of dimensions **m x n**, consisting of **m** rows and **n** columns. Each row is an aligned reference sequence and column denotes locations where a nucleotide, gap or dot occurs as shown in Table 2-1. As the database represents multiple sequence alignment of 16S rRNA genes, dots are used for padding before the start and after the end of a reference sequence depending on how the sequence was aligned against other sequences and therefore are not factored in any calculation, as they do not signify any information. Gaps however were accounted for, when calculating Shannon

entropy. To simplify calculations, ambiguous sequences that contained nucleotides other than A, T, C or G such as N were removed from the database.

Table 2-1: Database as a m x n matrix where rows are sequences and columns represent alignment positions

	C_1	C_2	C_3	C_4	C_5	C_6	…	C_{n-1}	C_n
R_1	.	A	T	-	A	T	…	G	.
R_2	T	A	G	C	A	A	…	G	.
R_3	.	.	A	C	T	A	…	T	.
.	.	.	.				…		
.	.	.	.				…		
R_m	A	T	T	A	A	C	…	A	.
SE	S_1	S_2	S_3	S_4	S_5	S_6	…	S_{n-1}	S_n

Shannon entropy was calculated on a per column basis:

- For a column $\boldsymbol{C_n}$ in the database matrix $\boldsymbol{M}$:
    1) For every row, check the nucleotide or gap that occurs in the row.
    2) Increment corresponding sum of the relevant nucleotide or gap.
    3) When all rows are iterated over, store the sum of each nucleotide and gap.

With this, Shannon entropy was calculated for every column in the following manner:

- For every column in the aligned database:
    1) The probability of each nucleotide or gap was calculated by dividing the number of occurrences of the specific nucleotide or gap over total number of occurrences of all nucleotides and gaps as shown below:

       $P(n_i)$ = Probability of nucleotide n_i = (number of n_i) / (total number of all nucleotides and gaps)

    2) The probability *$P(n_i)$* generated for each nucleotide was then multiplied with its natural log, *Ln ($P(n_i)$)*.

    3) Shannon entropy of a column was then calculated as the sum total of Shannon entropy of every nucleotide and gap in the following manner.

$$-\sum_{1}^{i} p(n_i)\, ln\, p(n_i)$$

Calculation of Shannon entropy for every column in this manner resulted in a single large vector for the database, with each location storing the Shannon entropy of the respective column. Given that every reference sequence in the database had a location for each of its nucleotides, the location of every nucleotide and the global database level Shannon entropy vector was then used to generate a per reference sequence Shannon entropy vector.

This was accomplished in the following manner:

- As every row $\boldsymbol{R_m}$ represented a reference sequence in the matrix $\boldsymbol{M}$, the calculation was performed by iterating over every column:
    1) A per reference Shannon entropy vector was generated by storing entropy value of a column $\boldsymbol{C_n}$ if a nucleotide was present at this location.
    2) Gaps were ignored in this process, as they do not form part of the sequence.

As an example, R_3 reference sequence had nucleotides at location/column C_3 – C_6 … C_{n-1} as shown in Table 1. The corresponding Shannon entropy vector was therefore S_3 – S_5, … S_{n-1}.

In this manner, the Shannon entropy vectors for every reference sequence were generated and stored in a database and was only needed to be performed once.

2.2.3 Taxonomic Annotation Process

The system flowchart is illustrated in Figure 2-1, where USEARCH alignments (Edgar, 2010) were used to reconstruct full alignments between query sequences and reference 16S rRNA gene sequences. This determined precisely where matches, mismatches and gaps occurred against a reference sequence. Relative

entropy was then calculated using the vectors developed for each reference sequence and finally each query read was scored.

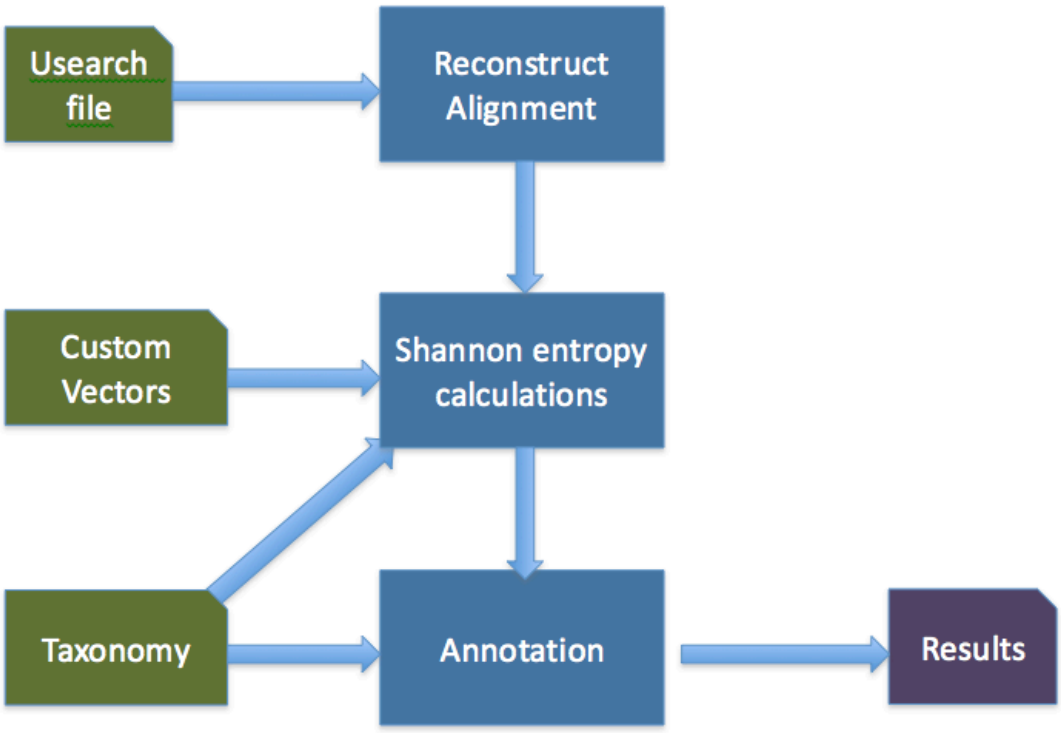


Figure 2–1: System process diagram where data files are shown in green, processing tasks in blue and results in purple

The process is described as below:

1) Query sequences were aligned with the reference SILVA database. The resultant data contained complete information of alignment between the reference and query sequences as well as the location of alignments.
2) Alignments were then reconstructed where location of gaps, matches and mismatches were determined.

3) Shannon entropy for each query sequence and the matched reference sequence segment was calculated using the stored vectors in a separate database.
4) Finally, relative Shannon entropy score was calculated and query sequences were annotated with reference sequence taxonomic annotation.

Shannon entropy for each alignment was then determined. Every reference sequence had a corresponding custom vector, or entropy vector. When an alignment occurred, the input read aligned at a certain region on the reference sequence.

The Shannon entropy for the query or input read was then calculated in the following manner:

- For an input read $\boldsymbol{I_i}$ having length $\boldsymbol{m}$ that aligned to a reference sequence $\boldsymbol{R_j}$:
    1) The location of matches between reference sequence and input read were found.
    2) For every match, the corresponding Shannon entropy value was taken from the database using the location of the match on the reference sequence.
        a. For example, if nucleotide $\boldsymbol{D_3}$ on input read matches nucleotide $\boldsymbol{Z_7}$ on reference sequence, the corresponding Shannon entropy value is $\boldsymbol{S_7}$.

Custom vector, V_j

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	…	S_n

Reference Sequence, R_j

Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	…	Z_n

Input Read, I_i

D_1	D_2	D_3	D_4	…	D_m

S_n: Entropy value at location n

Z_n: Reference sequence nucleotide Z at location n

D_m: Input read nucleotide D at location m

Next, reference read Shannon entropy was calculated using the segment of reference sequence included in the alignment.

2.2.4 Relative Shannon Entropy Score

For an input read of length m, reference read Shannon entropy is:

$$\sum_{i}^{j} S_i = \mathrm{SE}_{\mathrm{Ref}}$$

Where "i" is the location of start of alignment and "j" is the end of alignment. This denoted the total sum of entropy values from the start of alignment to the end of

alignment. As Shannon entropy was originally calculated using five variables (nucleotides A, T, C, G and "gap"), the maximum possible Shannon entropy is 1.609438 for any location.

Calculating input read Shannon entropy was done as follows:

$$\sum_{p}^{q} S_p = \mathrm{SE}_{\mathrm{Read}}$$

This denoted the total sum of entropy values for all the matches between the reference sequence and input read. When there was a complete matching between references read and input read, the total entropy value was the same between both. In the case of mismatches or gaps, the number of matches' decreases which leads to lower total Shannon entropy value for the input read.

Relative Shannon entropy for every input sequence was generated in the following manner:

1) Shannon entropy value on locations where a nucleotide mismatch occurred between the reference and query sequence was converted to a negative value for query sequence.
2) Next, for both reference sequence and query sequence, the maximum Shannon entropy value was added on each location. This enabled better segregation of sequences, which may contain mismatches, as the penalty of each mismatch as well as addition of each match is doubled and hence the overall score difference increases for closely related sequences.

3) Finally, the total entropy value for both the reference sequence segment as well as the query sequence was calculated by adding values at every location.
4) A relative entropy score was then calculated by dividing total Shannon entropy value of a query read by the total Shannon entropy value of the reference read segment. As every reference sequence had a taxonomic annotation associated with it, the matched input read was assigned this annotation.

$$Relative\ SE\ Score = {SE_{Read}}\Big/{SE_{Ref}}$$

The relative Shannon entropy score also acts as the scoring system, replacing percentage identity as the sequence similarity metric. The score was calculated using the formula below, with SE Coverage denoting the total number of bases in an alignment. This also factored gaps in the scoring as well.

$$Read\ Score = {SE_{Read}}\Big/{SE_{Ref}} * \frac{SE\ Coverage}{Alignment\ Length}$$

Once relative entropy had been scored for every input sequence, the results were stored in a text file. Furthermore, the system was based on a best scoring approach whereby the single best alignment was used for each input read. That is, the reference sequences with the best alignment score was used, instead of multiple alignments. This is in contrast to taxonomic annotation systems such as MEGAN, which use a least common ancestor algorithm whereby each read is assigned to

the lowest common ancestor of the set of taxa of reference sequences that the read hits in the comparison (Huson et al., 2007; Huson et al., 2009).

2.2.5 Validation

Validation of the system was performed using an *in-silico* approach. MicroSim: A motif-based next-generation read simulator developed by Schirmer *et. al.* was used to generate amplicon reads from reference sequences from SILVA database (Quast et al., 2013), simulating a motif-based illumina Miseq Fusion Golay V4 Amplicon 250bp (DS78) platform. The simulator provides a variety of profiles, targeting various amplicon and metagenomics based sequencing approaches. Amplicon sequencing was selected due to its prevalence in next-generation sequencing based taxonomic annotation projects.

Mock communities of sequences were generated using the SILVA release 123 database (Quast et al., 2013). The dataset was used to validate the Shannon entropy based annotation approach and confirm that it is producing similar taxonomic annotation (thus community composition) compared to percentage identity as calculated by USEARCH (Edgar, 2010). As the taxa assignments of the sequences selected to generate mock communities were already known, the following metrics were used in the validation process:

Recall: $\frac{TP}{TP+FN}$

Precision: $\frac{TP}{TP+FP}$

Accuracy: $\frac{TP+TN}{TP+FP+TN+FN}$

Where TP denotes True Positives, FP as False Positives, TN as True Negatives, and FN as False Negatives (Fawcett, 2006). These metrics are widely utilized for evaluation of classification systems and examples include studying association between heart failure self-management and rehospitalisation *via* natural language processing (Topaz et al., 2016), pre-miRNA precursor identification using neural networks (Jiang, Zhang, Xuan, & Zou, 2016) and automated quality assessment of radiologic interpretations (Hsu, Han, Arnold, Bui, & Enzmann, 2016).

In a classification system, precision is defined as the ratio of correctly labelled instances of a class that are retrieved, divided by total number of all instances that are labelled as members of the class. In essence it was the ratio of number of relevant or correctly assessed instances (True Positives) to the total number of irrelevant and relevant instances (True Positives and False Positives). Recall is the fraction of correctly labelled instances that are retrieved. In other words, it was the ratio of number of correctly labelled instances of a class retrieved (True Positives) to the total number of all instances belonging to that class (True Positives and False Negatives). Furthermore, a list of scoring thresholds was selected and at every threshold precision and recall were calculated for both percentage identity and the new Shannon entropy based metric. Considering that

typically 80% is considered as similarity at phylum level, hence an exponential list of thresholds was selected to elucidate the difference in sequence similarity between Shannon entropy and percentage identity based metrics. The list of threshold values is given in the Table 2-2.

Table 2-2: List of thresholds between 0 and 1 used for calculation of precision and recall

Thresholds (between 0 and 1)	
1	0.93
0.995	0.9
0.99	0.85
0.985	0.8
0.98	0.7
0.975	0.6
0.97	0.5
0.965	0.4
0.96	0.3
0.955	0.2
0.95	0.1
0.945	0.0
0.94	

2.2.6 Area Under The Curve

To perform a comparison between percentage identity and a Shannon entropy based approach, the area under the curve for Precision/Recall curve metric was used. A common method for classifier comparison, a higher *AUC* denotes better classification capabilities. In other words, between two classification systems, the classifier having a higher AUC is better performing than the other as it produces better classification results.

Additionally, the metric is effectively threshold independent as the graph only illustrates the precision and recall information. This is especially useful given that a wide variety of classification systems use different threshold levels for classification purposes (Fawcett, 2006). The list of precision and recall values for each approach was then used to calculate area under the curve using the trapz function in R (Tuszynski, 2014).

2.2.7 Generation of Whole SILVA Based Dataset

For initial system validation, 20,000 amplicon reads were generated via MicroSim using the whole of SILVA database version 123 (Quast et al., 2013). The reads generated were then aligned against the reference database using USEARCH sequence aligner (Edgar, 2010). Once the alignment results were generated, the taxonomic annotation system generates sequence annotation using the vector database file and Silva taxonomy. Percentage identity scores were generated by USEARCH (Edgar, 2010).

2.2.8 Generation of Removal of Whole Genera, Families and Class Datasets

Using the SILVA database, DNA data related to 100 genera, 10 families and 1 class were randomly selected and removed. Query sequences belonging to these removed taxa are effectively novel to the remaining sequences in the database and therefore should not closely match any of the taxa retained in the database. This approach can be useful in understanding how the system reacts to novel sequences that may present themselves in real datasets to which the database is naive (Lanzen et al., 2012).

The removed taxa are then used for random generation of sequences using MicroSim with the same parameters as for whole SILVA based dataset, namely motif-based illumina Miseq Fusion Golay V4 Amplicon 250bp (DS78) platform. This ensured consistency across all datasets. These newly generated sequences were then aligned against the reference Silva database that does not contain these sequences. Finally, analysis was done to assess how the sequences are being annotated and a precision vs. recall curve is generated for both Shannon entropy and percentage identity approaches.

The process followed for the generation of these sequences and annotations is detailed below:

Validation Approach:

1) 1 class, 10 families and 100 genera were randomly selected out of the database. It is important to note here that because taxa were removed randomly, hence another member of the level higher taxa may exist in

the database. For example, if genera G_1 and G_2 belong to the family F_1, and G_1 was randomly removed then genera G_2 may have existed in the database.

2) New reference databases were made which did not contain these selected class, families or genera. These class, families and genera were effectively "novel" sequences to these new databases.
3) From these selected taxa, 20000 amplicon reads were generated using MicroSim.
4) These amplicon reads were then aligned against the new reference databases.
5) Taxonomic annotation was performed. Here the approach was to check the immediate higher-level annotation and how many sequences annotated correctly at that level. For example, if the removed taxa was a family, then sequences were checked by testing how many were correctly annotated to the order level, additionally including sequences that were below order level as well (genus and family).

The list of genera, families and class removed from the SILVA release 123 database are given in Table 2-3.

Table 2-3: List of taxa removed at genus, family and class level from SILVA database

Genera Removed	**Families Removed**	**Class Removed**

Acetatifactor	Methyloligella	Acidaminococcaceae	Clostridia
Actinobacillus	Naasia	Actinomycetaceae	
Actinotignum	Neiella	CHAB-XI-27	
Adlercreutzia	Nisaea	Fervidicoccaceae	
Alkalitalea	Oceanobacter	Lactobacillaceae	
Amycolatopsis	Oceanobacterium	Lentisphaeraceae	
Arcobacter	Oleibacter	Microbacteriaceae	
Balnearium	Pasteuria	Peptostreptococcaceae	
Basfia	Planomicrobium	SM1B06	
Budvicia	Pleionea	nbr16a11	
Butyrivibrio	Pragia		
Caldisericum	Quadrisphaera		
Campylobacter	Rarobacter		
Chelonobacter	Rhizobacter		
Chroococcus	Rhizobium		
Collinsella	Rhodobium		
Cruoricaptor	Rhodonellum		
Delftia	Roseobacter		
Desulfomicrobium	Rothia		
Desulfurispira	Rudaea		
Eisenbergiella	Rudaibacter		
Enterobacter	Solobacterium		
Epibacterium	Spiribacter		
Eremococcus	Spongiibacter		

Ferroglobus	Spongiibacterium		
Flaviramulus	Telluria		
Flavivirga	Terrabacter		
Fluviicola	Terrimonas		
Frischella	Thauera		
Gaiella	Thermoflexus		
Gelria	Thermoproteus		
Geobacter	Thermus		
Gleocapsa	Thiofaba		
Haliea	Thioploca		
Halovenus	Thioreductor		
Hamadaea	Ureaplasma		
Hellea	Vadicella		
Ideonella	Vibrio		
Ignicoccus	Volucribacter		
Janibacter	Waddlia		
Jonesia	Wandonia		
Kineosphaera	Wenxinia		
Kineosporia	Woodsholea		
Leeia	Xenococcus		
Limnobacter	Xenophilus		
Lonsdalea	Yangia		
Malikia	Zavarzinia		
Mameliella	Zhouia		

Marinilabilia Zymobacter Methylobacter Methylobacterium		

2.2.9 Toolset Developed

The aforementioned tasks were accomplished by developing a collection of tools and scripts. The list of these tools and scripts alongside their description is listed in Table 2-4.

Table 2-4: List of tools and scripts

Script/Tool	**Description**
usearch_makeudb	This script converts FASTA files such as SILVA database to USEARCH UDB format for use in sequence alignment. SILVA reference database was broken down into smaller files due to memory limitations of 32-bit USEARCH aligner.
usearch_align	This script performs sequence alignment of datasets via USEARCH aligner with reference database in UDB format.

reduceusoutput.jar	Using the USEARCH generated results from each individual UDB file, this tool selects the alignments with highest percentage identity score and discards the rest.
SilvaUtils-*.jar	A set of tools that perform a variety of tasks on the SILVA release 123 database. From the fully aligned SILVA database, ambiguous reads and eukaryote sequences were removed, and suitable unaligned and aligned databases were generated. Additionally, RNA sequences were converted to DNA sequences here as well.
rdpse-s.jar	This tool uses a fully aligned SILVA database to generate entropy information for each reference sequence and stores it in a text file. The process is only needed to be done once.
uploadvectoS3db.jar	This tools stores entropy information generated for each reference sequence via the rdpse-s.jar tool to a

	SQLITE3 database, which can then be used in generating Shannon entropy based scores for query sequences. This significantly enhanced the throughput of the system.
TaxaSE.jar	The main Shannon entropy based taxonomic annotation system. The tool used a SQLITE3 database file containing entropy information for reference sequences as well as alignment results generated from reduceusoutput.jar. The system then outputs Shannon entropy based results.
Silva-slicer-*.jar	A set of tools that assists in removal of taxa based validation approach. The SILVA database is broken or "sliced" on the basis of genus, family and class level from which taxa can be removed and new databases can be generated.
singlethresholdvalidator.jar	This tool assists in the validation process. The tool consisted of two parts, which calculated precision and recall for whole SILVA dataset and

	removal of taxa based datasets respectively. The tool used a list of thresholds and iterated across them, calculated both precision and recall at each step. The Shannon entropy based results as well as the original FASTA File and SILVA taxonomy data are used to determine whether each query sequence is annotated correctly as per requirement.

2.3 Results

2.3.1 Whole SILVA Based Dataset

For the whole SILVA based dataset, the precision vs. recall curves for Shannon entropy and percentage identity approaches is illustrated in Figure 2-2. Both approaches demonstrated similar performance. Given that this was a simulation of an Illumina sequencing system, the precision varied between approximately 0.967 and 0.964.

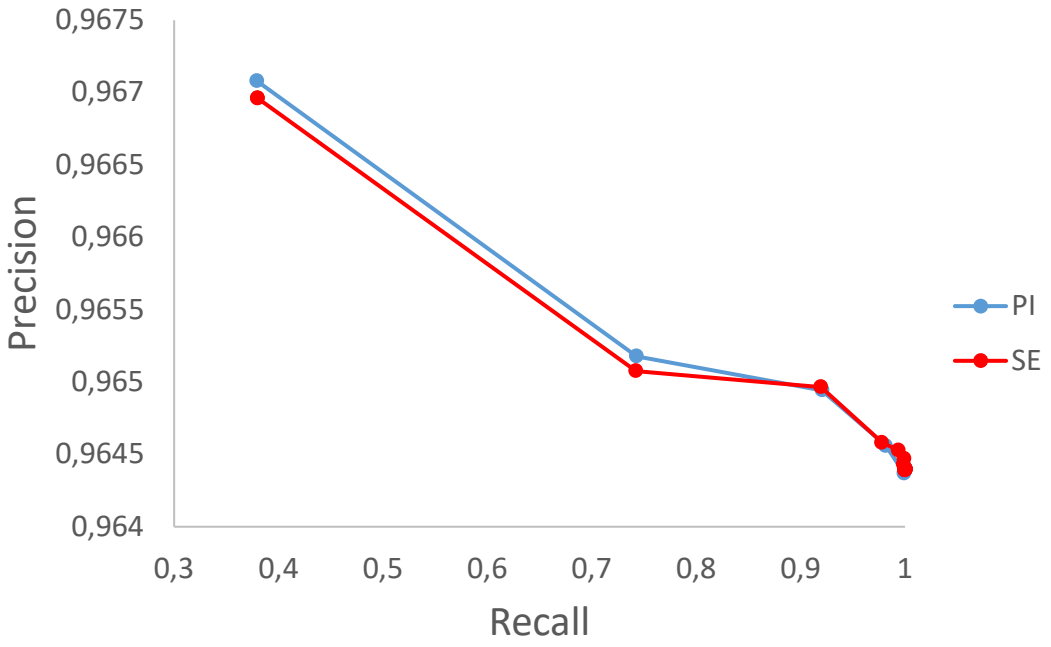


Figure 2–2: Precision vs. recall graph for whole SILVA dataset with percentage identity in blue and Shannon entropy approach in red

The area under the curve for both approaches is given in Table 2-4. Both approaches generated an area under the curve of 0.599, with percentage identity being slightly ahead of Shannon entropy based approach.

Table 2-5: Area under the curve for whole SILVA dataset based validation for both the percentage identity and Shannon entropy approach

Area Under the Curve	
Percentage identity	0.5992876
Shannon Entropy	0.5991252

2.3.2 Removal of Whole Genera, Families and Class

As with the aforementioned whole SILVA dataset based validation, the precision/curve of both Shannon entropy and percentage identity approaches closely match each other for removal of genera based dataset (Figure 2-3). Precision started at less than 0.5, diminishing as recall improved for both approaches.

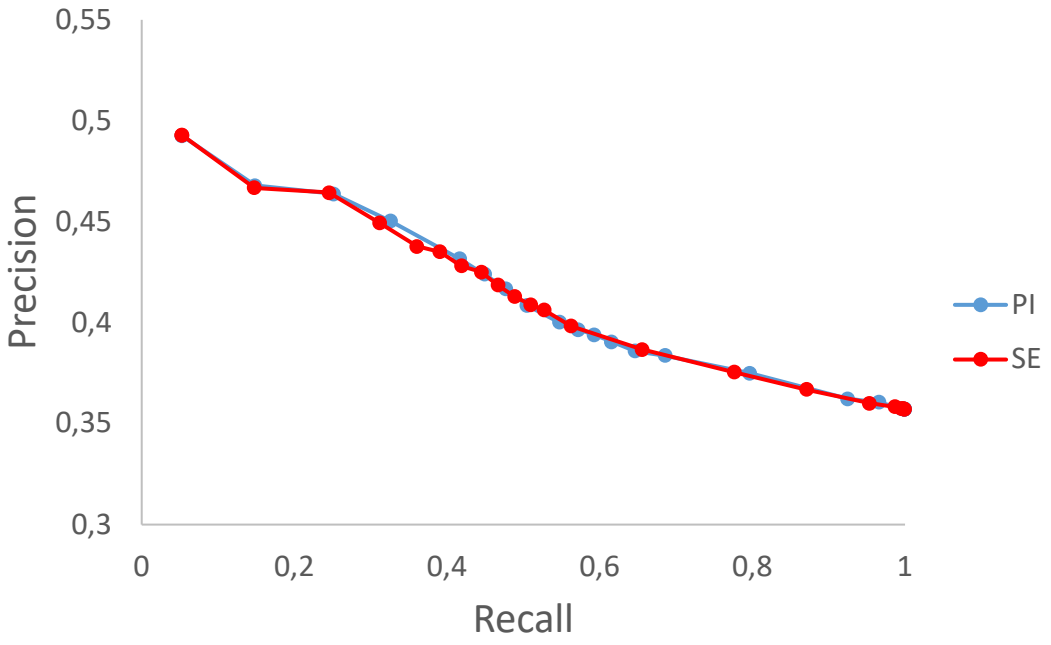


Figure 2–3: Precision vs. recall graph for removal of genera dataset with percentage identity in blue and Shannon entropy approach in red

The area under the curve calculated for both approaches are shown in Table 2-6. Removal of genera based validation showed percentage identity slightly outperforming our Shannon entropy based approach. Given that this was a test of both systems on novel sequences, precision was low overall, staying below 0.5.

Table 2-6: Area under the curve for removal of genera based validation for both percentage identity and Shannon entropy approach

Area Under the Curve for Removal of Genera based Validation	
Percentage identity	0.3924979
Shannon Entropy	0.3917333

The precision vs. recall curves for removal of families-based validation is illustrated in Figure 2-4. For most of the graph, the precision vs. recall curve for Shannon entropy stayed above the precision vs. recall curve for percentage identity. Precision for both curves began at 0.4 and stayed below this until full recall was achieved. Table 2-6 lists the area under the curve for both approaches.

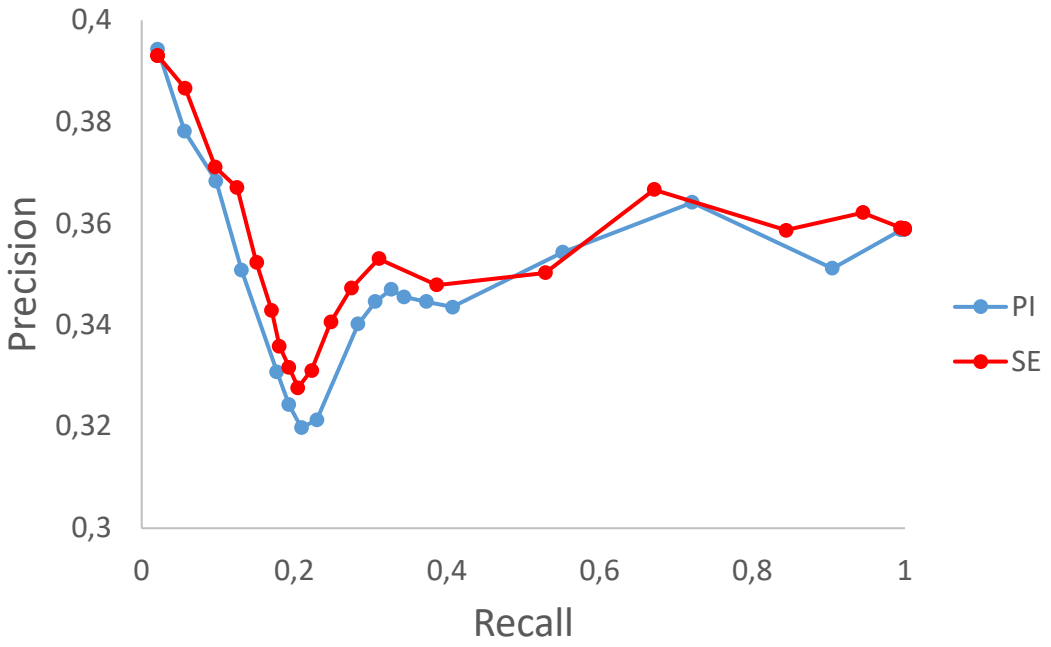


Figure 2–4: Precision vs. recall graph for removal of families dataset with percentage identity in blue and Shannon entropy approach in red

For the removal of families based validation, both approaches generated similar area under the curve as shown in Table 2-7, however Shannon entropy achieved a slightly higher AUC of 0.349 while percentage identity attained 0.345.

Table 2-7: Area under the curve for removal of families based validation for both percentage identity and Shannon entropy approach

Area Under the Curve for Removal of Families based validation	
Percentage identity	0.3448928
Shannon Entropy	0.3493627

The precision vs. recall curves for removal of class-based validation approach is shown in Figure 2-5. Precision was low for both approaches, staying below 0.4. Furthermore, the areas under the curve for both approaches followed each other closely.

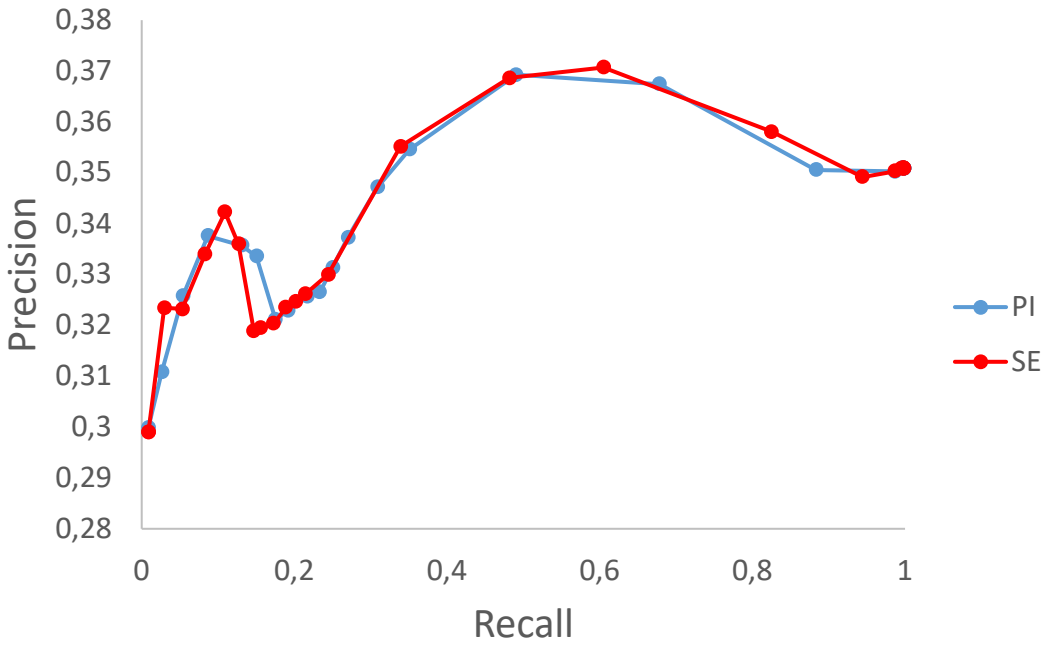


Figure 2–5: Precision vs. recall graph for removal of class dataset with percentage identity in blue and Shannon entropy in red

The calculated area under the curve for both approaches showed a slight advantage towards Shannon entropy approach, which scored an AUC of 0.348 while percentage identity achieved 0.347 (Table 2-8).

Table 2-8: Area under the curve for removal of class based validation for both percentage identity and Shannon entropy approach

Area Under the Curve for Removal of Class based validation	
Percentage identity	0.347263
Shannon Entropy	0.3478349

2.3.3 Accuracy

Accuracy plots for both approaches were generated by varying the thresholds and calculating accuracy attained at each threshold. The results were then plotted as a graph between calculated accuracy and the threshold used. Both the whole SILVA and removal of taxon-based datasets were used.

The graph of accuracy against thresholds for whole SILVA dataset is illustrated in Figure 2-6. Accuracy for percentage identity rose quickly with earlier thresholds compared to Shannon entropy, where lower thresholds are needed to attain similar accuracy. At the threshold value of 0.975, both approaches achieve the

maximum accuracy possible and the graph hits a plateau afterwards, keeping the accuracy unchanged.

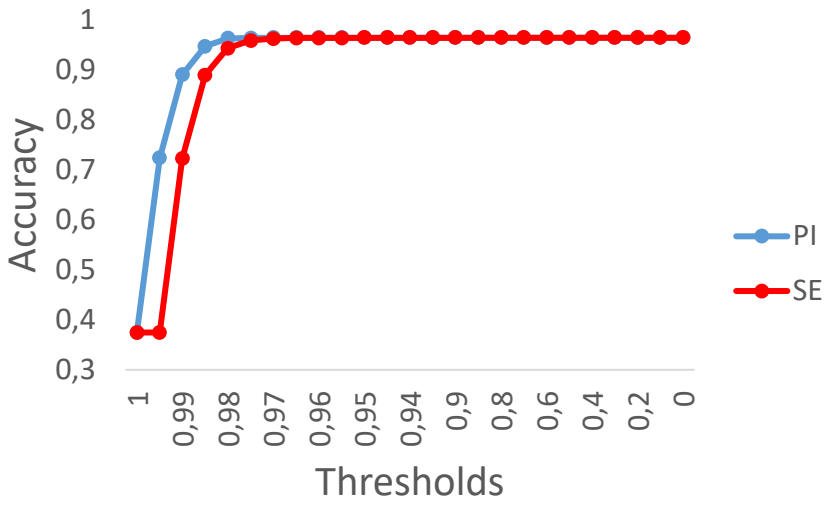


Figure 2–6: Whole SILVA dataset accuracy graph for percentage identity in blue and Shannon entropy in red

Accuracy graphs for the removal of taxa based datasets are illustrated in Figure 2-7. Similar to whole SILVA accuracy graph, here as well the thresholds were varied and accuracy was calculated at every threshold. Percentage identity demonstrated the similar behaviour as before, where accuracy rose quickly with each threshold while reaching the maximum value at 0.7 thresholds for all three validation approaches (removal of genus, families and class).

Shannon entropy followed the same pattern with lower threshold needed for equivalent accuracy. Additionally, the accuracy value for the highest two thresholds, 1 and 0.995 (denoting 100% and 99.5% sequence similarity respectively) were the same.

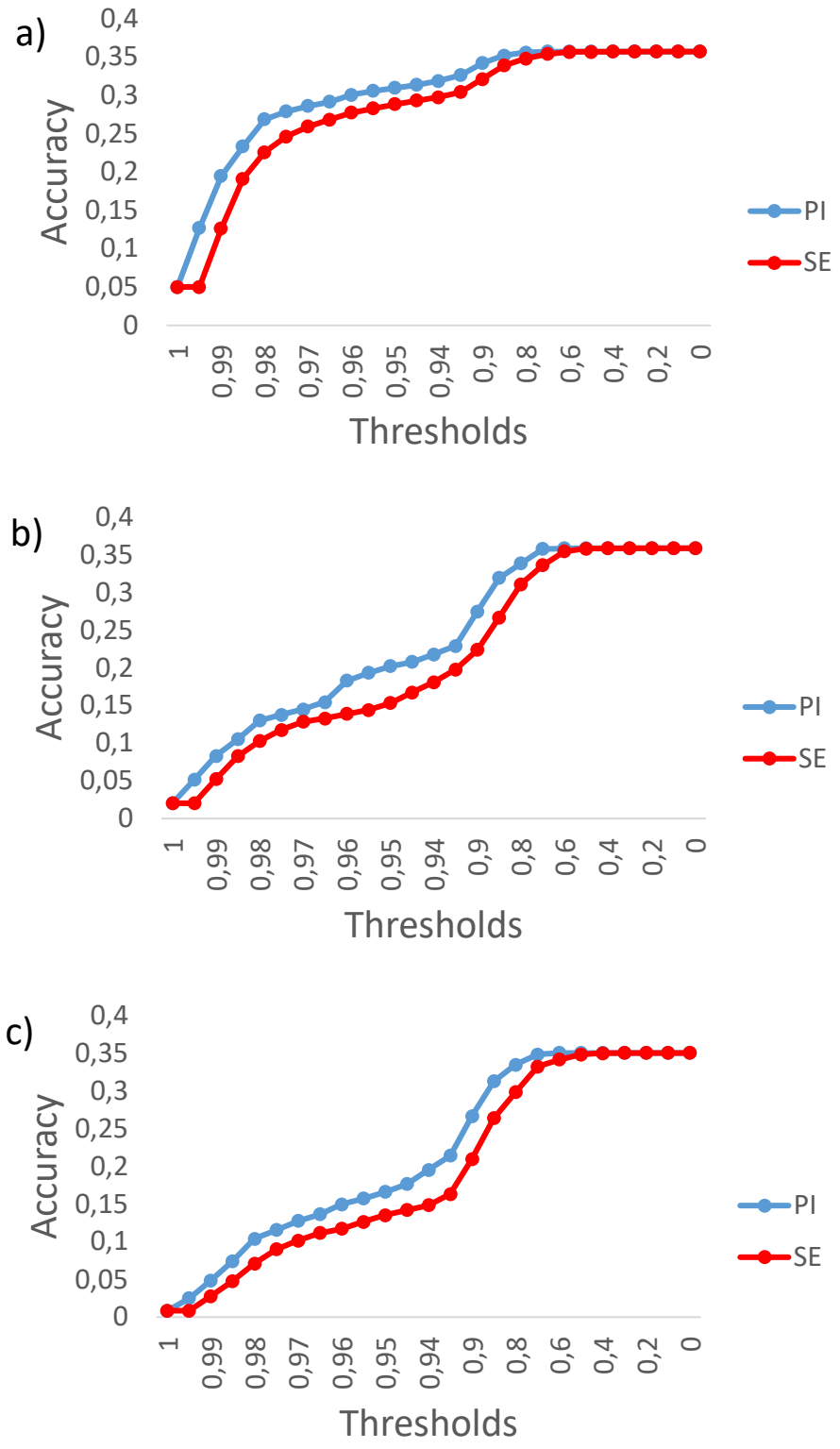


Figure 2–7: Removal of taxa accuracy graphs with percentage identity in blue and Shannon entropy in red at three taxa levels, a) genus, b) family and c) class

2.4 Discussion

The new Shannon entropy based sequence similarity metric can be used as a replacement of industry standard percentage identity. The new approach showed comparative performance for whole SILVA dataset and slightly lower for removal of genus validation dataset. However, it improved upon percentage identity for removal of families and classes datasets.

2.4.1 Whole Silva Dataset

Shannon entropy approach closely follows the Percentage identity based approach and therefore shows similar performance to the gold standard percentage identity. Given the similar performance, using either approach would most likely to produce almost same results when tested on datasets containing previously described 16S rRNA gene sequences.

2.4.2 Removal of Taxa Dataset

For removal of genus dataset, sequences were checked at family level. Both approaches generated almost the exact same result in this case, with percentage identity slightly leading over Shannon entropy approach.

However, the Shannon entropy based approach showed improved performance compared to percentage identity based approach, with higher AUC in the case of

removal of families dataset. For removal of class dataset, sequences were checked at phylum level and while both approaches were similar in their capability, Shannon entropy based approach demonstrated slightly improved performance. This translates into better annotation of novel sequences at the order level as well as phylum level compared to percentage identity based approach and is therefore much more effective at taxonomic annotation as novel sequences can be annotated better in the case of the new approach. By randomly selecting the taxon to be removed at each taxa level in an independent fashion, the AUC results of both approaches are a good indicator of their performance as it minimized any possible bias towards specific taxa groups. Later chapters in this study where more samples and replicates are included also illustrate an improvement as well. Lastly, the different behavior shown by the precision vs recall curves for removal of taxa may be due to a few factors, such as inaccurate taxonomic annotation in the reference database, low quality sequences or presence of polyphyletic sequences.

A low accuracy was achieved in the removal of taxon datasets for both approaches. This is due to the fact that these sequences are effectively novel to the database and hence a lower accuracy is expected. It is possible that popular classifiers such as RDP and SILVA may also have similarly lower accuracy, as sequence similarity based classifiers tend to perform worse when the query sequence is distantly related to the sequences in the database (Matsen et al., 2010). Additionally, any threshold used for percentage identity approach (Lanzen et al., 2012) may not be appropriate for Shannon entropy approach, due to the fundamental difference in the scoring of query sequences. The Shannon entropy approach scores differently compared to percentage identity, as the mismatch penalty depends on the location

of the mismatch and the associated entropy value and therefore achieved similar accuracy to percentage identity at a later threshold (Figure 2-6 and 2-7). Additionally, the increase in accuracy with decreasing threshold differs for different levels of taxa.

Percentage identity is a distance-based approach where only the number of positions where sequences differ, is used to calculate a similarity score. Because various segments of 16S rRNA gene are changing at different rate, evolutionary distances between sequences are not captured effectively in this manner and in fact are underestimation (Woese, 1987). Furthermore, it also suffers from lower taxonomic resolution especially at the species level, as sequences belonging to different species can be erroneously considered identical as only the number of mismatches is counted but not the locations where these mismatches occur (Fox et al., 1992). Unlike the aforementioned percentage identity, the new Shannon entropy based approach effectively captures evolutionary conservation from the 16S rRNA gene sequences as every location's degree of variability is directly determined and used in the new scoring scheme. This represents advancement towards better similarity measurements and which is in accordance with the evolution of sequences (Woese, 1987). The results illustrate better annotation capability at class and families level while being comparative to percentage identity at other taxa levels. A limitation of the new method is that only the best alignment was taken. The impact on the results may be small as it is more likely that the same taxonomic annotation is assigned to the query sequences as the whole SILVA dataset illustrates a very high precision, even with simulated error in reads. Nonetheless, it is possible that the application of least common ancestor

on multiple alignments may yield better results, as this would allow for differentiating between assignments that are very closely related.

An alternative to quantifying evolutionary conservation via Shannon Entropy would be the application of Stochastic Context Free Grammar based tools such as SSU-align (Nawrocki, Kolbe, & Eddy, 2009). While keeping the model structure contact, the emission and transition probabilities can be trained for each taxonomic level, which is then used for developing multiple models. A query sequence could then be searched against all models, with the best model being selected. Another avenue would be to use SCFG based tools for generating better multiple sequence alignment of reference sequences (Brown, 2000) to more accurately determine Shannon entropy at each location. Lastly, for metagenomic datasets, SCFG based homology search may outperform sequence similarity based tools such as BLAST (Yuan, Lei, Cole, & Sun, 2015).

Given that the vast majority of sequences are uncultivated (Huson et al., 2007; Marcy et al., 2007), there is a higher likelihood that in many ecological studies unknown sequences will be detected. The best possible annotation of these sequences will give insight into the inner workings of the environment, even if the exact taxonomic annotation cannot be determined at finer taxonomic levels (Huson et al., 2007). For this reason, new approaches must be able to handle these sequences in an improved fashion and here the new Shannon entropy based approach provides improved performance over the industry standard percentage identity, by annotating novel sequences better at higher taxa levels such as family

and class. This illustrates the advantage of using Shannon entropy approach over percentage identity metric.

Additionally, techniques such as oligotyping (Eren, Borisy, Huse, & Mark Welch, 2014) have already utilized Shannon entropy albeit for a few hundred base pairs. Shannon entropy is calculated across the whole of 16S rRNA gene in the new approach, enabling it to capture variability in the gene much more effectively. Furthermore, using a fast sequence aligner in the form of USEARCH enables quick and high throughput taxonomic annotation where large datasets can be quickly annotated in short time (Edgar, 2010).

2.5 Conclusion

The study aimed to develop a novel entropy based approach that can replace the percentage identity metric for sequence similarity where the evolutionary conservation information of 16S rRNA genes are directly exploited to form a new high resolution scoring method. Most popular approaches forgo the utilization of this inherent information contained within the 16S rRNA sequences, instead relying on a measure that only counts mismatches between sequences. Given the variability across the whole of 16S rRNA, not every base may be equally important as variable locations are much more essential in differentiating between sequences compared to conserved regions (Chakravorty et al., 2007).

The approach is competitive enough that it can be used alongside commonly applied percentage identity scoring schemes. The new approach performs slightly worse for whole SILVA and removal of genera based validation, although the performance is just within reach of percentage identity. Furthermore, Shannon entropy based approach shows improved performance for removal of families and class based validation. This is especially important given that majority of bacterial sequences are not annotated, and more and more novel sequences are being detected in almost all of the next-generation sequencing projects. Hence new approaches, which are able to annotate novel sequences at various taxa levels, would be more appropriate and Shannon Entropy based approach may be more suitable for this purpose.

The SILVA database used contained more than 1.2 million fully aligned sequences. Increase in the number of aligned sequences belonging to a wide variety of diverse bacteria can improve the capability of the system, as full alignment of 16S rRNA gene sequences are the central focus of the new approach described here and an increase in the diversity present in the aligned database would improve the system due to generation of Shannon entropy information that captures variability in the 16S rRNA marker gene much more effectively. Higher quality sequences as well as species level information would no doubt lead to better taxonomic annotation as well.

Chapter 3: TaxaSE: Taxonomic Annotation via Shannon Entropy

3.1 Introduction

Advances in sequencing technology have led to an explosion in the amount of biological data being generated (Thorsen et al., 2016). Given the importance played by microbes in the inner working of the environment (Kirk et al., 2004; Mackelprang et al., 2016) and the effects on human health (Stefka et al., 2014; von Mutius, 2016), numerous studies are being conducted to elucidate the various mechanisms by which these microbes influence their surroundings (Gilbert, Jansson, & Knight, 2014). As a consequence, bioinformatics pipelines aiming to characterize microbial community composition, have been developed alongside various 16S rRNA gene sequence databases, which serve as a reference set of sequences for microbial taxonomic analysis (Santamaria et al., 2012).

Popular taxonomic annotation pipelines include MG-RAST (Aziz et al., 2008), MEGAN (Huson et al., 2007), QIIME (Caporaso et al., 2010) and MOTHUR (Schloss et al., 2009). MG-RAST (Aziz et al., 2008) is an online service for phylogenetic and functional annotation of metagenomes. MEGAN (Huson et al., 2007) is a standalone tool, primarily geared towards taxonomic annotation of metagenomes while QIIME (Caporaso et al., 2010) and MOTHUR (Schloss et al., 2009) are suites of bioinformatics tools, which provide a flexible workflow for analysis of microbial communities.

QIIME or Quantitative Insights into Microbial Ecology (Caporaso et al., 2010) is a popular pipeline package used in various ecological projects. It packages together commonly used algorithms such as USEARCH, a sequence aligner (Edgar, 2010), BLAST (Altschul et al., 1990), Denoiser (Reeder & Knight, 2010) or AmpliconNoise (Quince et al., 2009) for denoising reads, Uchime (Edgar et al., 2011) for chimera removal and also includes sequence handling and statistical tools. Various studies that have used QIIME for analysis include studies of the structure, function and diversity of the human microbiome (Human Microbiome Project, 2012), gut microbiota (Claesson et al., 2012; Turroni et al., 2012), soil bacterial communities (N. Fierer et al., 2012; Nacke et al., 2011; J. Rousk et al., 2010) and marine microbiota (Mason et al., 2012; Zettler, Mincer, & Amaral-Zettler, 2013). For most analysis, QIIME mainly uses the UCLUST (Edgar, 2010) algorithm for clustering and RDP classifier (Cole et al., 2014) for taxonomic assignment purposes.

The majority of taxonomic annotation systems, including QIIME, use OTU or operational taxonomic unit, as the defining concept for determining community composition (He et al., 2015). Considered as a *de facto* standard approach to analysis, OTUs are formed by clustering sequences on the basis of a specified similarity threshold such as 97% (Drancourt et al., 2000; Tikhonov, Leach, & Wingreen, 2015). Taxonomic annotation is performed on the representative sequence of each OTU, and all the sequences within the OTU are assigned the same taxonomy regardless of small scale differences in base composition between them (Nguyen, Warnow, Pop, & White, 2016). This is a favorable technique as picking representative OTUs from a list of sequences drastically cuts down on

computational requirements for analysis. This gives the ability to quickly perform fast annotation, in addition to providing abundance information of how many reads form an OTU cluster (He et al., 2015; Methé et al., 2012) and therefore allowed for rapid analysis of large datasets (Nguyen et al., 2016).

However, both taxonomic annotation and OTU generation suffers from various limitations. Being a non-evolutionary based distance metric (Nguyen et al., 2016), the percentage identity metric, which is used to determine sequence similarity for both taxonomic annotation and OTU generation, provides an inaccurate estimation of evolutionary distance between two sequences (Woese, 1987). As only the number of mismatches is used to calculate the percentage identity metric, taxonomic annotation at species level also suffers, with the approach unable to differentiate between closely related species or strains (Fox et al., 1992).

OTU generation methods assume that all 16S rRNA genes evolve at the same rate (Schloss & Westcott, 2011) as they are dependent on the simple distance based metric like percentage identity and hence cannot capture the variability at each nucleotide position. Furthermore, OTUs made from short read sequences may not be as reliable in estimating species richness as the OTUs formed from near full-length sequences, primarily due to 16S rRNA gene exhibiting different degree of variability across its length and therefore region selection plays an important role in accurately estimating microbial diversity (Minseok Kim, Morrison, & Yu, 2011).

Additionally, OTU assignments may not be reliable and can differ on the basis of the algorithm used (Tikhonov et al., 2015), with common OTU creation

approaches sometimes leading to inflation of species level diversity estimates (Edgar, 2013; White et al., 2010). This is compounded by the fact that certain OTU construction techniques generate instable OTUs where the membership of sequences changes significantly with the addition of new sequences or samples to the dataset. As a consequence, different set of OTUs are observed with each clustering run (He et al., 2015). Sequences belonging to one OTU in the previous run may be assigned to different OTUs in the next run and sequences belonging to different OTUs may be merged into a single OTU. This has a significant impact on downstream diversity analysis including rarefaction curves, which determine how much diversity was captured as well as the identification of individual OTUs (He et al., 2015; Nguyen et al., 2016). Given that most taxonomic annotation pipelines regularly employ percentage identity based OTU clustering for any downstream analysis, these pipelines therefore suffer from the same limitations as well. However, newer tools such as PhylOTU use phylogenetic distances instead of percentage identity to generate OTUs, though these are targeted towards shotgun metagenomics datasets (Sharpton et al., 2011).

This study aims to address these issues and overcome these limitations by developing a new taxonomic annotation pipeline, defined here as Taxonomic Annotation *via* Shannon entropy (the TaxaSE system), which employs the novel Shannon entropy based sequence similarity measure as developed in Chapter 2. As the Shannon entropy based approach exhibits better performance compared to percentage identity in some instances, based on Chapter 2's *in-silico* analysis on algorithm performance, therefore this would result in better taxonomic annotation capability compared to percentage identity based approaches. While

the new pipeline can be used for annotation of OTUs, the limitations associated with OTU generation and usage can be resolved by following an OTU-independent approach where sequences are annotated individually. This would result in the highest resolution annotation via a combination of an improved annotation algorithm as well as extracting intra-OTU diversity, compared to standard 97% OTU similarity approach, which obscures fine-scale variation. This requires more computational resources but the system is able to attain species level annotation. Hence, this chapter aimed to develop a taxonomic annotation pipeline, which could use the newly developed sequence similarity metric described in chapter 2 and to test the pipeline on real amplicon datasets and in comparison to another popular pipeline, QIIME.

To perform an exploration of TaxaSE's capabilities towards annotation of diverse microbial sequences and its value for application in ecological studies, comparison with a published and widely applied pipeline is needed. For that reason, QIIME (Caporaso et al., 2010) is selected due to its popularity in ecological studies and because it provides various tools for downstream analysis of taxonomic annotations and is underpinned by similarity based annotations *via* the popular USEARCH sequence aligner (Edgar, 2010). Furthermore, TaxaSE was integrated into the QIIME workflow by specially developing tools that can convert the generated results into a QIIME compatible format. This will greatly enhance the ability to compare both pipelines as well as integrating QIIME based tools and the TaxaSE system. This integration has long-term benefits as the TaxaSE system can be deployed quickly in ecological studies where QIIME is already being used.

To demonstrate the usefulness of this approach, we applied it to amplicon datasets from specific habitats in order to generate meaningful results and capture diversity patterns in a similar manner to QIIME.

3.2 Materials and Methods

3.2.1 Pipeline Development

The TaxaSE pipeline was developed in the Java programming language and represented a collection of tools and scripts developed for taxonomic annotation and integration with QIIME. Description of the tools and scripts developed and the various tasks they performed are listed in Table 3-1. The pipeline's development was based on handling different aspects of analysis by modules of scripts and tools, and the workflow used is as follows:

1) Sequence alignment: This uses USEARCH sequence aligner to generate alignments between reference and query sequences. As the free version of USEARCH aligner is 32bit and therefore has a memory limitation, hence the SILVA database was broken down into more manageable parts and scripts and Java tools were developed for using multiple reference files with USEARCH and generating the best read.
2) Shannon Entropy read score generator: This was developed to use best sequence alignments to generate a Shannon entropy score for each query sequence. The results consisted of query sequences and their associated Shannon entropy read scores for use in downstream tools.
3) Threshold based conversion: This was developed to select the maximum taxonomic annotation level of query sequences on the basis of thresholds. The result file here consists of query sequences and the final taxonomic annotations assigned to them.

4) QIIME conversion: Lastly, this enabled the conversion of TaxaSE results in a flat file to QIIME compatible format, allowing the use of TaxaSE within QIIME.

The pipeline toolkit, source code, associated dataset and documentations on how to run it for analysis is available publicly at HIE-Pub (Ijaz, 2017).

Table 3-1: Lists of tools and scripts developed for the TaxaSE pipeline.

Script/Tool	**Description**
usearch_makeudb	This script converts SILVA database to USEARCH UDB format for use in sequence alignment. SILVA reference database was broken down into smaller files due to memory limitations of 32-bit USEARCH aligner.
usearch_align	This script performs sequence alignment of datasets via USEARCH aligner with reference database in UDB format
reduceusoutput.jar	USEARCH generated results from each individual UDB file. This tool selects the alignment with highest percentage identity score and discards the rest.

TaxaSE.jar	The main Shannon entropy based taxonomic annotation system. The tool used a SQLITE3 database file containing entropy information for reference sequences as well as alignment results generated from reduceusoutput.jar. The system then outputs Shannon entropy based results.
se_threshold_converter.jar	This tool used a list of thresholds in a text to convert annotation to proper level in the results generated from TaxaSE.jar tool.
se_to_qiime.jar	Using the information present in TaxaSE results file, this tool generated QIIME compatible files, which can then be used within QIIME, pipeline itself for analysis.

3.2.1 Sampling

Dr. Kelly Hamonts at HIE, Western Sydney University, collected sugarcane leaf, stalk, root and rhizosphere soil samples in November 2014 from eight sugarcane fields growing three sugarcane varieties (KQ228, MQ239 and Q240) near Ingham,

Queensland, Australia. In each field, 3 stools were randomly selected and samples were collected from 2 plants per stool. Samples were snap-frozen in liquid nitrogen in the field, transported to the laboratory on dry ice and stored at -80 ^{o}C. Frozen sugarcane tissue samples were ground using mortar and pestle and DNA was extracted from the resulting powder using the MoBio PowerPlant DNA extraction kit, following the manufacturer's instructions. The MoBIO PowerSoil DNA extraction kit was used to extract DNA from the soil samples. Bacterial 16S rRNA amplicon sequencing was performed by the NGS facility at Western Sydney University using Illumina Miseq (2x 301 bp PE) and the 341F/805R primer set for this study.

A total of 158 samples, belonging to these environments were analyzed for comparison between TaxaSE and the RDP classifier, a naïve Bayesian classifier (Cole et al., 2014), as implemented in QIIME (Caporaso et al., 2010). The breakdown of the samples from the sugarcane dataset is listed in Table 3-2. Most samples came from the soil environment, followed by stem and root, while the rhizosphere environment had the least number of samples.

Table 3-2: Sample data used for real amplicon dataset analysis

Environment	**Number of Samples**
Rhizosphere	12
Root	45
Soil	54
Stem	47

Total	158

To minimize DNA sequencing artifacts and remove chimeras from the datasets, the following preprocessing procedure was followed for all samples:

1) Read trimming:
    a. Sequences were trimmed on both forward R1 and reverse R2 reads removing low quality regions with Phred (Ewing, Hillier, Wendl, & Green, 1998) score of less than 25 (Q25). This was performed using "seqtk" tool (Li, 2017).
2) Paired-end read merging:
    a. After quality trimming, both forward and reverse reads were merged using FLASH (Magoc & Salzberg, 2011) with a maximum overlap set to 200.
3) Chimera removal:
    a. Finally, the merged reads were analyzed for the presence of chimeras. This was accomplished using VSEARCH, a sequence aligner and RDP (Cole et al., 2014) Gold database which contained 10,049 reference sequences. Subsequently, chimeras were removed from the samples.

3.2.2 Comparison Approaches

In order to properly compare the new TaxaSE system with QIIME, the following two approaches were taken:

1) *OTU Comparison:* OTUs were generated from each of the four habitats at 97% sequence similarity using UCLUST (Edgar, 2010). Both QIIME and TaxaSE systems were run on the representative sequences of these OTUs and diversity results were compared. This was done to determine whether TaxaSE could be used as an integrated 16S rRNA gene sequence annotator for OTU based analysis.
    a. Sample files belonging to different environments were combined and a list of OTUs were generated using UCLUST (Edgar, 2010). This ensured consistency, as similar OTUs could then be compared.
    b. OTU tables were rarified to 10000 sequences to ensure even sampling depth.
    c. RDP classifier (Cole et al., 2005) was used for QIIME, with a default confidence parameter of 0.8.
    d. Given that representative sequences were based on the 97% OTU similarity, a single threshold of 0.9 was selected for TaxaSE on an ad hoc basis, as it is sufficiently higher for annotation of OTUs. This is similar to how RDP classifier in QIIME using a single parameter of 0.8 for annotation purposes. Furthermore, the selection of 97% OTU similarity means that the representative sequence to be annotated was already highly similar to other sequences in the OTU.
2) *Distinct Taxonomic Annotations:* Instead of OTUs, the number of distinct taxonomic annotations was used as the metric of diversity. To illustrate the comparison between annotating sequences individually with OTU based approaches, the following steps were taken:

a. OTUs were generated at 97% and 99% sequence similarity using QIIME. Following the annotation process via RDP classifier, OTUs, which had the same taxonomic annotations, were combined together to form pseudo-OTUs.
b. For TaxaSE, sequences were individually annotated and similar to QIIME, collections of sequences were combined together on the basis of having the same taxonomic annotations to form pseudo-OTUs.
c. The OTU tables were rarefied to 10000 sequences to ensure even sampling depth.

Furthermore, the following steps were also taken to ensure consistency between both approaches:

1) QIIME specific SILVA (Pruesse et al., 2007) database v119 was used for QIIME based analysis. For the TaxaSE system, SILVA database v123 was utilized for annotation purposes.
2) OTUs belonging to eukaryota and archaea were removed from QIIME results as the primary comparison between both systems was based on bacterial taxonomic annotations.

Given that sequences were annotated individually in the distinct taxonomic annotation approach, a set of thresholds was selected for the TaxaSE system, which is listed in Table 3-3. As from Chapter 2, the Shannon entropy based similarity metric reached similar accuracy at a lower threshold compared to percentage identity. Furthermore, while the percentage identity metric

underestimates evolutionary distances between sequences, the Shannon entropy based approach provided a more accurate assessment. Hence the threshold selected for each taxa level were selected on an ad hoc basis and were slightly lower than the corresponding thresholds for percentage identity, which are generally taken as 99% for species, 97% for genus, 95% for family, 90% for order, 85% for class and 80% sequence similarity for phylum (Drancourt et al., 2000; Lanzen et al., 2012).

Table 3-3: Thresholds selected for the TaxaSE system at different taxa levels

Thresholds	**Assigned Taxonomic Level**
1 – 0.98	Species or best possible annotation
0.98 – 0.95	Genus
0.95 – 0.9	Family
0.9 – 0.85	Order
0.85 – 0.8	Class
0.8 – 0.75	Phylum

3.2.3 Diversity Analysis

For the aforementioned two approaches, the following diversity analysis were performed:

1) Alpha diversity analysis was implemented using QIIME's inbuilt *alpha_rarefaction.py* script.
    a. Alpha diversity is used to quantitatively analyze the species richness in a habitat.
    b. Welch's t-test was used to determine if the results were statistically significant.
2) Beta diversity analysis was accomplished by using QIIME's *beta_diversity_through_plots.py* script. Bray Curtis was taken as the distance metric and beta diversity plots were generated using the Emperor package (Vázquez-Baeza, Pirrung, Gonzalez, & Knight, 2013).
    a. Beta diversity is a comparison of diversity between ecosystems and is therefore a useful step for comparing different taxonomic annotation pipelines.
3) Quantitative comparison of both QIIME and TaxaSE pipelines was performed via ADONIS (Anderson, 2001) and ANOSIM (CLARKE, 1993) statistical tests. The *compare_categories.py* script was used for this purpose.
    a. ADONIS represents a non-parametric multivariate analysis of variance, which uses distance matrices such as the Bray Curtis metric and details how much variance is described by a categorical variable.

b. ANOSIM represents analysis of similarities to test whether group of samples is statistically different based on a categorical variable.

3.3 Results

3.3.1 OTU Comparison

3.3.1.1 Alpha Diversity

For OTU comparison, the rhizosphere environment showed the TaxaSE system, using Shannon entropy at a single threshold of 0.9 producing the greatest number of OTUs at 3502 average while QIIME generated 3482 average as illustrated in Figure 3-1a. Standard error for observed species was the least in the case of QIIME with 59.6 while TaxaSE generated a standard error of 63.5. Welch's t-test was conducted to compare both approaches, where no significant difference was observed between QIIME (M=3482.09, SD=197.8) and TaxaSE (M=3502.3, SD=210.77); t (19)=0.2316, p = 0.8193.

In the case of root environment, both approaches provided almost the same results as shown in Figure 3-1b. QIIME generated a slightly higher 3452 average number of observed species, followed by TaxaSE at 3448. The standard error for observed species was also similar, with QIIME at 78.5 and TaxaSE at 79. Welch's t-test reported no significant difference between QIIME (M=3452.4, SD=496.7) and TaxaSE (M=3448, SD=499.9); t (77) = 0.0393, p = 0.9688.

For the soil environment, TaxaSE slightly lagged behind QIIME and generated an average observed species at 3811 as shown in Figure 3-1c, while QIIME generated 3816. Observed species standard error for TaxaSE system was at 80.5, while QIIME produced 79.4. Welch's t-test results showed no significant difference

between QIIME (M=3816.02, SD=572.47) and TaxaSE (M=3811.75, SD=580.6); t (101) = 0.0378, p = 0.97.

Lastly for the stem environment, TaxaSE provided 2071 as the average observed species, followed by QIIME at 2064 (Figure 3-1d). Standard error was similar across all approaches, with QIIME at 56.3 while TaxaSE produced a standard error of 56.8. Finally, Welch's t-test reported no significant difference between QIIME (M=2064.8, SD=382.07) and TaxaSE (M=2071.3, SD=385.4); t (89) = 0.0815, p = 0.9352.

Overall, TaxaSE system provided taxonomic annotation for that largest number of OTUs in two of the four environments, namely rhizosphere and stem. For root and soil, the new pipeline followed QIIME closely, which annotated higher number of annotated OTUs. As no statistically significant differences were found for all environments between both TaxaSE and QIIME, therefore the new pipeline performed at a similar level to QIIME for annotation of OTUs.

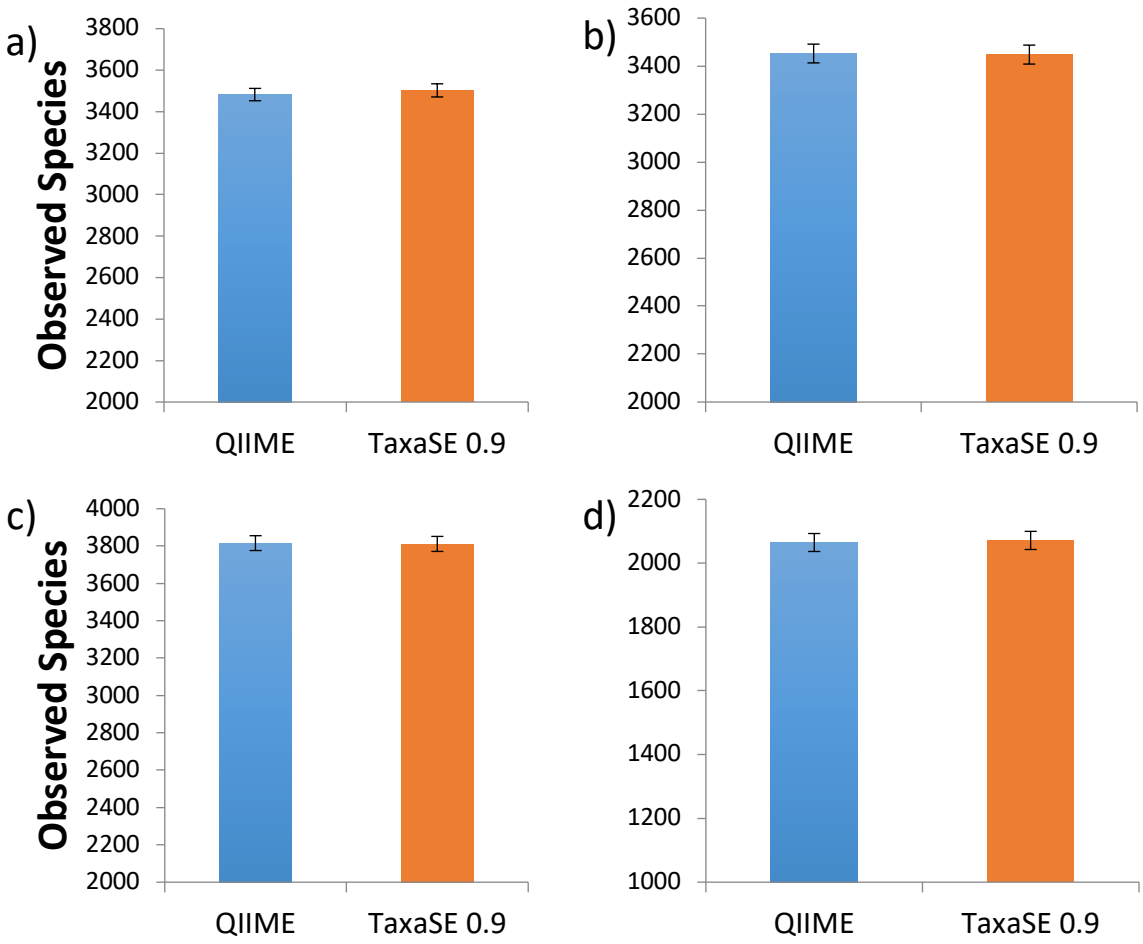


Figure 3–1: Observed species for OTU comparison at 97% OTU similarity with a) rhizosphere, b) root, c) soil and d) stem. QIIME is shown in blue while TaxaSE is shown in orange. Error bars represents standard error.

Shannon diversity results for rhizosphere environment showed similar patterns between QIIME at 97% OTU similarity and TaxaSE system as illustrated in Figure 3-2a. TaxaSE scored a Shannon diversity average of 10.039, followed closely by QIIME at 10.03. Standard error for Shannon diversity was also similar, where TaxaSE and QIIME scored standard error at 0.1. Welch's t-test reported no

significant difference between QIIME (M=10.03, SD=0.333) and TaxaSE (M=10.04, SD=0.334), t (19) = 0.0678, p = 0.9466.

Root environment samples exhibited similar Shannon diversity results between both pipelines as well, with an average value of 9.61 as shown in Figure 3-2b. The standard error observed for Shannon diversity came up slightly higher for TaxaSE at 0.126, while QIIME generated a standard error of 0.125. No statistically significant difference was found by Welch's t-test between QIIME (M=9.613, SD=0.792) and TaxaSE (M=9.612, SD=0.795), t (77) = 0.0069, p = 0.9945.

Soil environment results follow the same pattern as observed for rhizosphere and root environments. TaxaSE produced an average Shannon index of 10.26 while QIIME generated an index of 10.27 as displayed in Figure 3-2c. Standard error was slightly higher in the case of TaxaSE at 0.098, with QIIME coming up at 0.097. Welch's t-test reported no significant difference between QIIME (M=10.27, SD=0.7) and TaxaSE (M=10.26, SD=0.708), t (101) = 0.0725, p = 0.9423.

Finally, for the stem environment, TaxaSE generated a Shannon diversity index of 6.35 while QIIME produced 6.34 (Figure 3-2d). Standard error was also very similar with TaxaSE producing a standard error of 0.136 while QIIME generated 0.135. Here as well, Welch's t-test showed no statistically significant difference between QIIME (M=6.342, SD=0.917) and TaxaSE (M=6.35, SD=0.92), t (89) = 0.0463, p = 0.9632.

Overall, Shannon diversity results for all four environments illustrated similar behavior between QIIME and the TaxaSE system, with comparable index values and error rates observed.

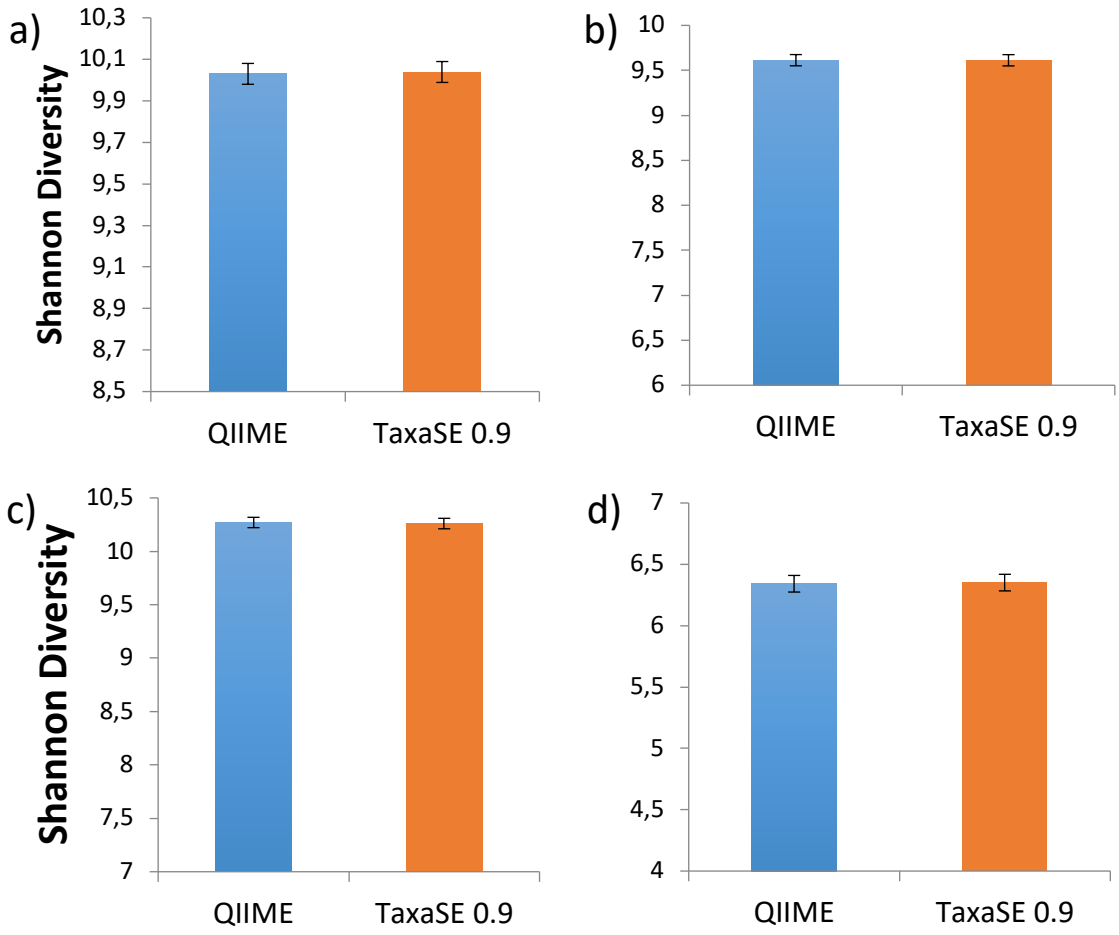


Figure 3–2: Shannon diversity for OTU comparison at 97% OTU similarity with a) rhizosphere, b) root, c) soil and d) stem. QIIME is shown in blue while TaxaSE is shown in orange. Error bars represent standard error.

TaxaSE performed in a similar manner to QIIME based approach and generated comparable results, where TaxaSE annotates slighter higher number of OTUs for rhizosphere and stem environments. Overall, no statistically significant

differences were found between TaxaSE and QIIME and hence the new pipeline captured similar patterns as the QIIME/Uclust method and therefore can be effective in generating alpha diversity analysis. Soil environment had the most OTUs as well as a higher Shannon diversity index, followed by rhizosphere and root environments. In contrast to these, Stem significantly showed less diversity, with a remarkably less number of OTUs generated as well as a lower Shannon diversity index.

3.3.1.2 Beta Diversity

The beta diversity plot for TaxaSE using 97% OTU similarity is illustrated in Figure 3-3a. Samples from stem were segregated from the rest of the habitats, while samples from root and soil were also mostly distinct from each other. The first principle coordinate axis, PC1, explained 30.64% of the variability.

The beta diversity plot for QIIME at 97% OTU similarity is shown in Figure 3-3b. Similar to the TaxaSE beta diversity plot, the samples from stem environment are distinct from the rest of habitats with root and soil samples showing some segregation as well. Here as well, the first axis is able to explain 30.64% of variability, in the same manner as TaxaSE system.

Both approaches showed a similar segregation of samples on the basis of environment, where the segregation pattern were similar across both approaches, and PC1, PC2 and PC3 for both approaches explained a similar amount of

variability. With respect to beta diversity analysis, TaxaSE system provided similar ecological patterns as were generated by QIIME.

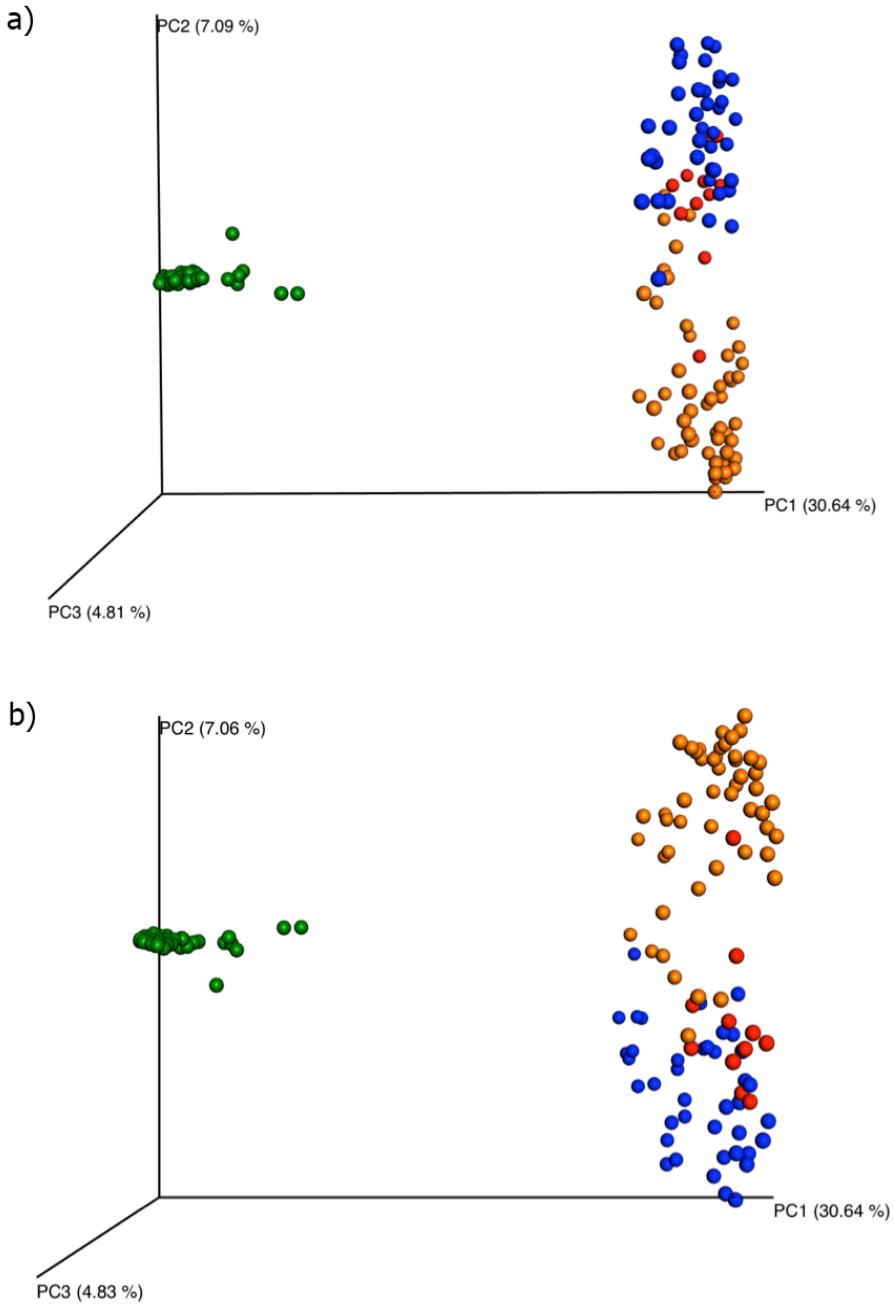


Figure 3–3: Beta diversity principle coordinate analysis plots for OTU comparison of sugarcane dataset with a) TaxaSE and b) QIIME. Rhizosphere samples are shown in red, root in blue, soil in orange and stem in green.

3.3.1.3 ADONIS and ANOSIM

ADONIS and ANOSIM tests were conducted to determine how much variation both QIIME and TaxaSE system explained, with the results listed in Table 3-4 for ADONIS. TaxaSE system captured the same variation when samples are grouped by habitats as by QIIME and therefore is able to capture similar patterns. Here TaxaSE produced a R^2 value of 0.37767 while QIIME produced a R^2 value of 0.3776.

Table 3-4: ADONIS results for OTU comparison at 97% similarity between TaxaSE and QIIME

TaxaSE at 0.9						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	20.354	6.7846	29.331	0.37767	0.001
Residuals	145	33.540	0.2313		0.62233	
Total	148	53.894			1.00000	
QIIME						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	20.352	6.7841	29.323	0.3776	0.001
Residuals	145	33.547	0.2314		0.6224	
Total	148	53.899			1.0000	

ANOSIM results for both pipelines show that grouping of samples on the basis of environments was strong for both methods, with R-value that was close to +1 (Table 3-5). Here TaxaSE generated an R-value of 0.855 while QIIME produced an R-value of 0.8553.

Table 3-5: ANOSIM results for OTU comparison at 97% similarity between TaxaSE and QIIME

Approach	p-value	R-value
TaxaSE at 0.9	0.001	0.855043
QIIME	0.001	0.855334

3.3.2 Distinct Taxonomic Annotations

3.3.2.1 Alpha Diversity

The alpha rarefaction plots (Figure 3-4) for all three approaches show that the TaxaSE system produced a higher number of observed species across all four environments as compared to both QIIME at 97% OTU similarity and QIIME at 99% OTU similarity.

Samples belonging to stem environments were less diverse than samples from rhizosphere, root and soil environments based on the number of observed species, which were far fewer.

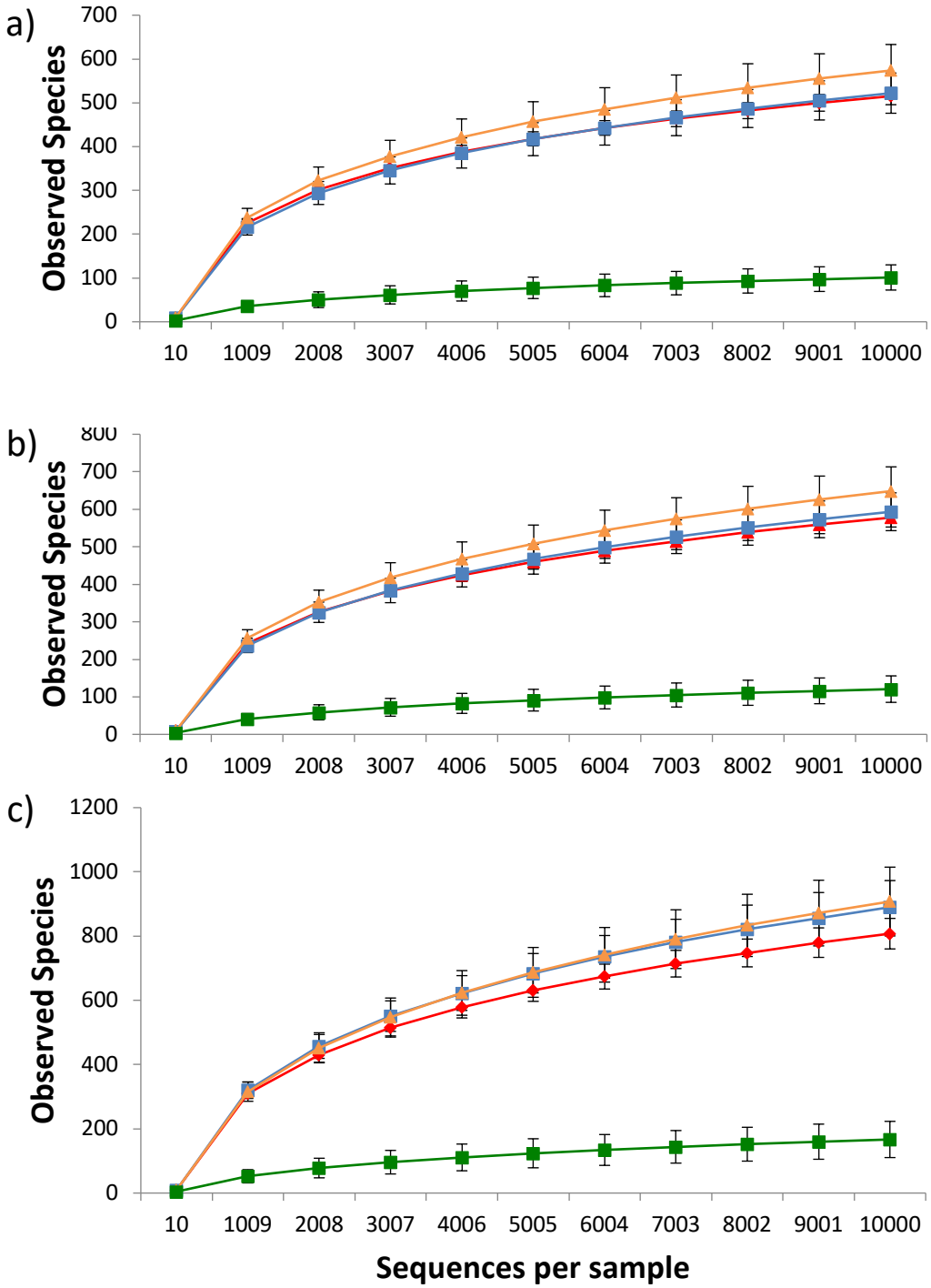


Figure 3–4: Alpha rarefaction plots for distinct taxonomic annotations of sugarcane dataset using a) QIIME at 97%, b) QIIME at 99% and c) TaxaSE. Rhizosphere samples are shown in red, root as blue, soil as orange and stem as green. Error bars represent standard deviation.

For rhizosphere environment samples, TaxaSE produced the highest number of distinct taxonomic annotations at 807, while QIIME at 99% OTU similarity produced 578 distinct taxonomic annotations and QIIME at 97% OTU similarity coming up last at about 515 as illustrated in Figure 3-5a. The standard error observed was highest in the case of TaxaSE system with 30.05, with QIIME at 99% OTU similarity at 15.89 and lastly QIIME at 97% OTU similarity at 12.3.

Welch's t-test showed a statistically significant difference between QIIME at 97% OTU similarity (M=515.18, SD=40.79) and QIIME at 99% OTU similarity (M=577.91, SD=51.71), t (18) = 3.1216, p = 0.0059. Furthermore, Welch's t-test also reported a statistically significant difference between QIIME at 97% OTU similarity (M=515.18, SD=40.79) and TaxaSE (M=807.64, SD=99.68), t (13) = 9.0059), p = 0.0001. Lastly, the difference was also statistically significant between QIIME at 99% OTU similarity (M=577.91, SD=51.71) and TaxaSE (M=807.64, SD=99.68), t (15) = 6.7572, p = 0.0001. All three approaches were therefore statistically different from each other, with TaxaSE pipeline generating the largest number of annotations.

For the root environment, here as well TaxaSE produced the largest number of distinct taxonomic annotations at 890, followed by QIIME at 99% OTU similarity with 593 distinct annotations and lastly QIIME at 97% OTU similarity at 522 (Figure 3-5b). Standard error for TaxaSE stood at 26.56, while QIIME at 99% had a standard error of 16.33 and lastly QIIME at 97% had 14.89.

Welch's t-test illustrated a statistically significant difference between QIIME at 97% OTU similarity (M=522, SD=92.96) and QIIME at 99% OTU similarity (M=593.41, SD=102), t (75) = 3.2315, p = 0.0018. A statistically significant difference was observed *via* Welch's t-test between QIIME at 97% OTU similarity (M=522, SD=92.96) and TaxaSE (M=890.08, SD=167.99), t (61) = 12.0882, p = 0.0001. Finally, the difference was also statistically significant between QIIME at 99% OTU similarity (M=593.41, SD=102) and TaxaSE (M=890.08, SD=167.99), t (64) = 9.514, p = 0.0001. Similar to the results for rhizosphere environment, TaxaSE pipeline produced the highest number of statistically significant distinct taxonomic annotations.

Soil showed similar pattern as with previous environments, with TaxaSE generating higher number of distinct taxonomic annotations reaching 907, while QIIME at 99% OTU similarity followed it at 697 annotations and QIIME at 97% OTU similarity coming up last at 574 distinct annotations (Figure 3-5c). TaxaSE system had the highest standard error at 29.99, while QIIME at 99% OTU similarity was at 18.1 and lastly QIIME at 97% OTU similarity at 16.6.

A statistically significant difference was observed *via* Welch's t-test between QIIME at 97% OTU similarity (M=573.75, SD=119.67) and QIIME at 99% OTU similarity (M=648.52, SD=130.54), t (101) = 3.0445, p = 0.003. A statistically significant difference was observed between QIIME at 97% OTU similarity (M=573.75, SD=119.67) and TaxaSE (M=907.67, SD=216.23), t (79) = 9.7433, p = 0.0001. Finally, the difference was also statistically significant between QIIME at 99% OTU similarity (M=648.52, SD=130.54) and TaxaSE (M=907.67, SD=216.23),

t (83) = 7.3987, p = 0.0001. Hence, similar to aforementioned environments, soil habitat produced a clear lead for TaxaSE pipeline.

Stem was the least diverse of all habitats, and while TaxaSE generated a higher number of distinct taxonomic annotations at 167, it also produced the highest standard error as well, at 17.82 (Figure 3-5d). QIIME at 99% OTU similarity generated 121 distinct annotations with a standard error of 11.03 while QIIME at 97% OTU similarity produced 101 distinct annotations with a standard error of 9.004.

The difference was not statistically significant, as found by Welch's t-test between QIIME at 97% OTU similarity (M=101.19, SD=58.35) and QIIME at 99% OTU similarity (M=120.71, SD=71.48), t (78) = 1.3713, p = 0.1742. However, a statistically significant difference was found between QIIME at 97% OTU similarity (M=101.19, SD=58.35) and TaxaSE (M=166.88, SD=114.08), t (59) = 3.2905, p = 0.0017. The difference was statistically significant between QIIME at 99% OTU similarity (M=120.71, SD=71.48) and TaxaSE (M=166.88, SD=114.08), t (66) = 2.2031, p = 0.0311. In contrast to previous environments, QIIME at 97% OTU similarity and QIIME at 99% OTU similarity performed in a similar manner here, although the new TaxaSE pipeline produced more distinct taxonomic annotations while being statistically different from either of these two approaches.

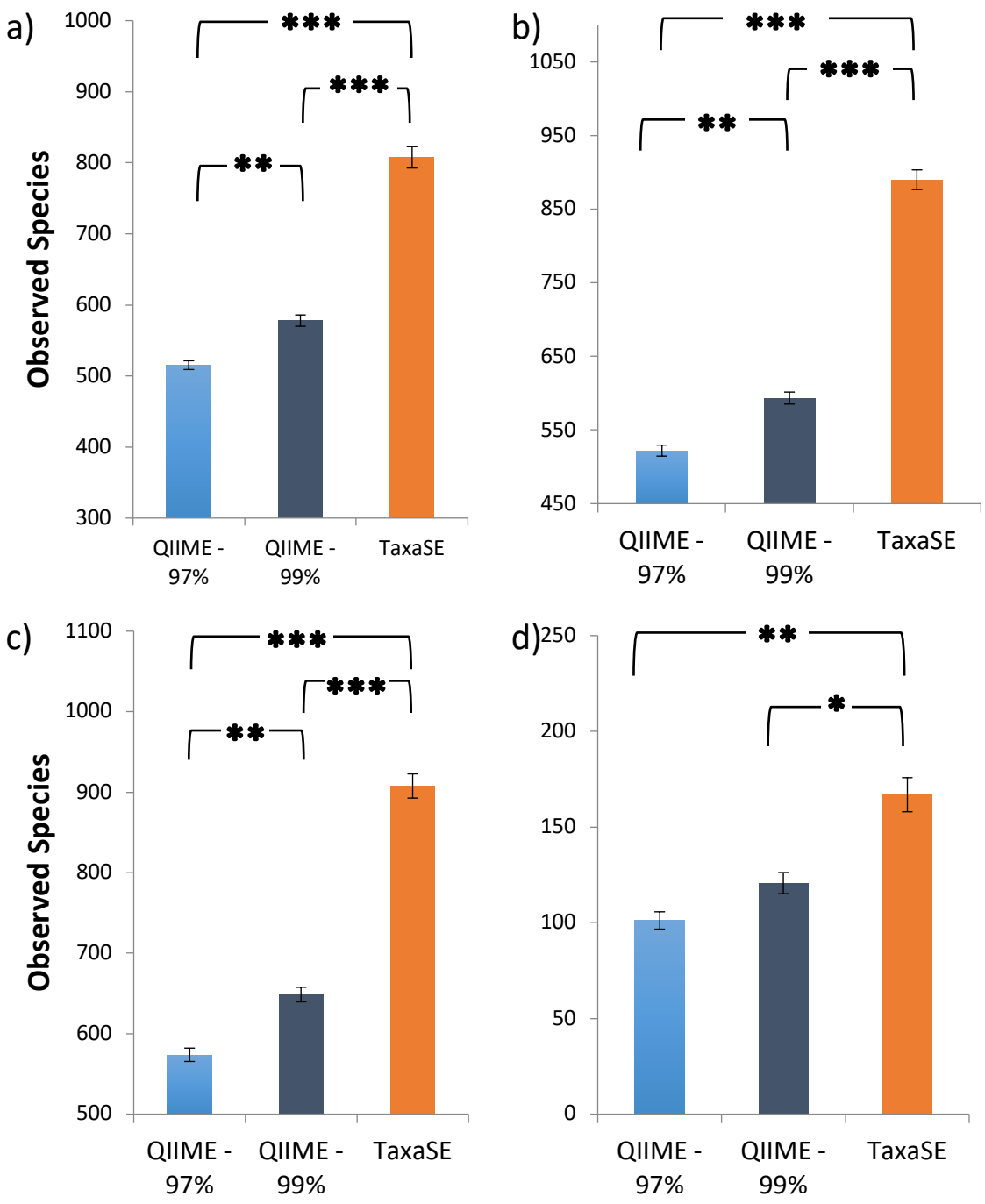


Figure 3–5: Observed species for distinct taxonomic annotation comparison with a) rhizosphere, b) root, c) soil and d) stem. QIIME at 97% OTU similarity is shown in blue, QIIME at 99% OTU similarity in dark blue and TaxaSE in orange. Error bars represent standard error. Significance levels are shown with asterisks, where * represents $p < 0.05$, ** represents $p < 0.01$ and * represents $p < 0.001$.**

For rhizosphere samples, TaxaSE produced the highest Shannon diversity index for distinct taxonomic annotation based comparison, with a value of 7.7, shown in Figure 3-6a. Furthermore, QIIME at 99% OTU similarity produced a Shannon diversity index of 7.1 while QIIME at 97% OTU similarity produced 6.9 as Shannon diversity index. Standard error for TaxaSE was 0.084, with QIIME at 99% OTU similarity at 0.049 and QIIME at 97% OTU similarity at 0.052.

Welch's t-test produced a statistically significant difference between QIIME at 97% OTU similarity (M=6.94, SD=0.173) and QIIME at 99% OTU similarity (M=7.1, SD=0.163), t (19) = 2.146, p = 0.045. The difference was statistically significant between QIIME at 97% OTU similarity (M=6.94, SD=0.173) and TaxaSE (M=7.73, SD=0.277), t (16) = 7.9947, p = 0.0001. Lastly, the difference was also statistically significant between QIIME at 99% OTU similarity (M=7.1, SD=0.163) and TaxaSE (M=7.73, SD=0.277), t (16) = 6.5353, p = 0.0001.

Samples from the root environment showed similar Shannon diversity index results between the two QIIME methods (Figure 3-6b), with TaxaSE leading with more than 7.6, followed by QIIME at 99% OTU similarity with 6.8 and lastly QIIME at 97% OTU similarity at 6.6. Standard error observed was highest in the case of QIIME at 97% OTU similarity with 0.104, followed by QIIME at 99% OTU similarity at 0.099 and lastly TaxaSE at 0.093.

The difference was not statistically significant between QIIME at 97% OTU similarity (M=6.628, SD=0.648) and QIIME at 99% OTU similarity (M=6.829, SD=0.617), t (75) = 1.4059, p = 0.1639. However, the difference was statistically significant between QIIME at 97% OTU similarity (M=6.628, SD=0.648) and TaxaSE (M=7.673, SD=0.59), t (75) = 7.4996, p = 0.0001. Finally, Welch's t-test reported a statistically significant difference between QIIME at 99% OTU similarity (M=6.829, SD=0.617) and TaxaSE (M=7.673, SD=0.59), t (76) = 6.2192, p = 0.0001.

TaxaSE also had higher Shannon diversity results for soil samples compared to QIIME at 97% and QIIME at 99% (Figure 3-6c), where TaxaSE showed higher diversity index at 7.77 than both QIIME methods, with QIIME at 97% OTU similarity at 7.1 and QIIME at 99% OTU similarity at 7.3. The standard error observed were 0.078 for TaxaSE, 0.067 for QIIME at 99% OTU similarity and 0.069 for QIIME at 97% OTU similarity.

Welch's t-test illustrated that the difference was not statistically significant between QIIME at 97% OTU similarity (M=7.08, SD=0.499) and QIIME at 99% OTU similarity (M=7.266, SD=0.485), t (101) = 1.9296, p = 0.0565. However, the difference was statistically significant between QIIME at 97% OTU similarity (M=7.08, SD=0.499) and TaxaSE (M=7.771, SD=0.559), t (100) = 6.6509, p = 0.0001. Lastly, the difference was also statistically significant between QIIME at 99% OTU similarity (M=7.266, SD=0.485) and TaxaSE (M=7.771, SD=0.559), t (99) = 4.922, p = 0.0001.

Finally, Shannon diversity index results for all three methods for stem samples showed TaxaSE having an average Shannon diversity of 2.7 while QIIME at 99% OTU similarity produced 2.4 and finally QIIME at 97% OTU similarity produced the lowest Shannon diversity at 1.7 (Figure 3-6d). The standard error observed were 0.131 for TaxaSE, 0.01 for QIIME at 99% OTU similarity and 0.104 for QIIME at 97% OTU similarity.

The difference was statistically significant between QIIME at 97% OTU similarity (M=1.663, SD=0.676) and QIIME at 99% OTU similarity (M=2.411, SD=0.648), t (81) = 5.1809, p = 0.0001. Similarity, the difference was also found to be statistically significant between QIIME at 97% OTU similarity (M=1.663, SD=0.676) and TaxaSE (M=2.727, SD=0.839), t (76) = 6.3544, p = 0.0001. However, the difference was not statistically significant between QIIME at 99% OTU similarity (M=2.411, SD=0.648) and TaxaSE (M=2.727, SD=0.839), t (75) = 1.9165, p = 0.0591.

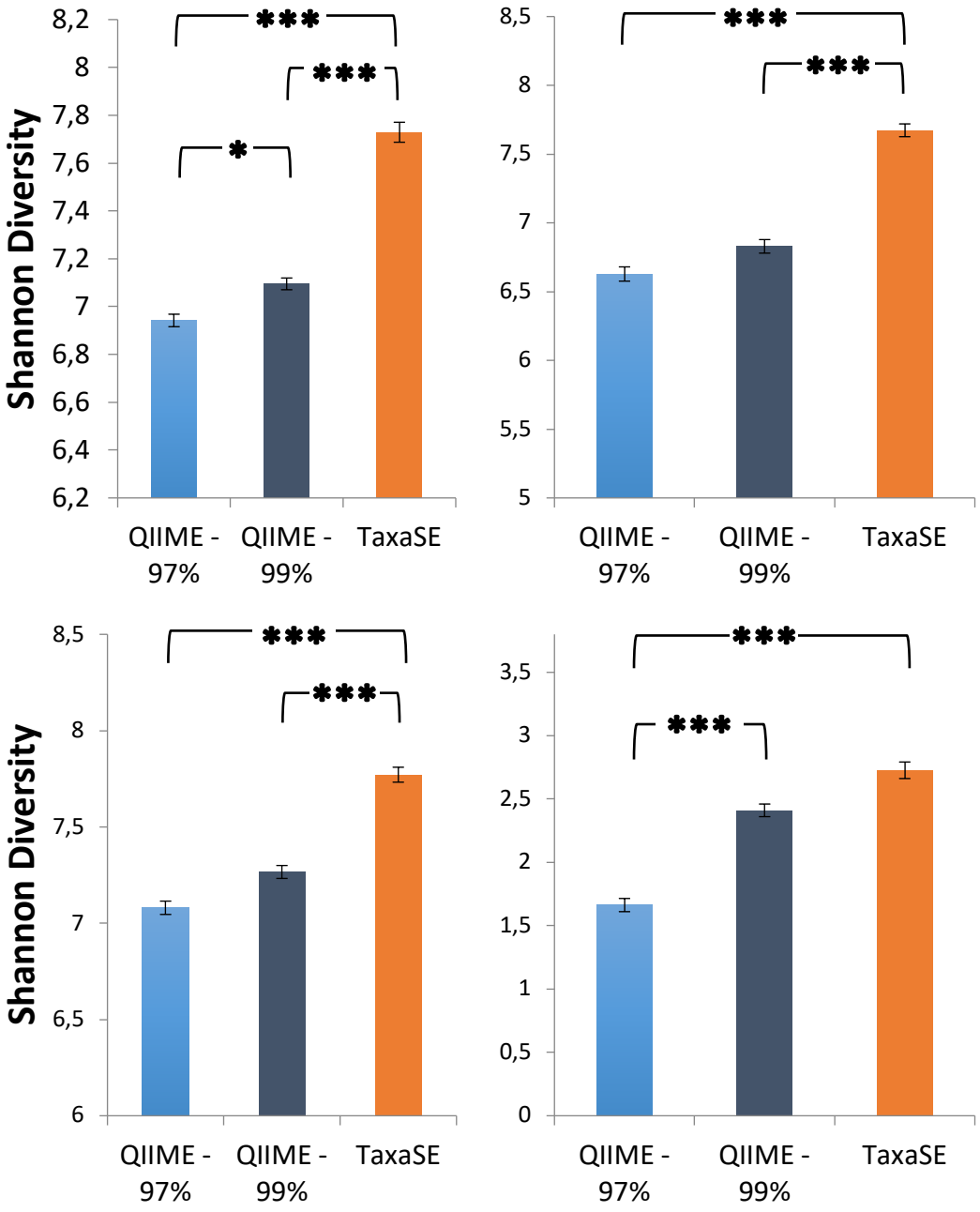


Figure 3–6: Shannon diversity for distinct taxonomic annotation comparison with a) rhizosphere, b) root, c) soil and d) stem. QIIME at 97% OTU similarity is shown in blue, QIIME at 99% OTU similarity in dark blue and TaxaSE in orange. Error bars represent standard error. Significance levels are shown with asterisks, where * represents $p < 0.05$, ** represents $p < 0.01$ and * represents $p < 0.001$.**

3.3.2.2 Beta Diversity

The beta diversity plot for QIIME at 97% OTU similarity is shown in Figure 3-7a. Stem samples were segregated from the samples belonging to other environments. Furthermore, root and soil samples displayed some segregation as well. The first principle coordinate, PC1 explained a variance of 58.31% in the case of QIIME at 97% OTU similarity.

Beta diversity plot for QIIME at 99% OTU similarity, as illustrated in Figure 3-7b, provided a similar pattern as was seen for QIIME at 97% OTU similarity (Figure 3-7a). Stem samples were segregated from the other samples and the first principle coordinate explained a variance of 57%, slightly lower than what was observed for QIIME at 97% OTU similarity.

Finally, the beta diversity plot for the TaxaSE system is shown in Figure 3-7c and here as well, stem samples were well segregated from other samples. Furthermore, soil samples were more densely packed along the first axis for TaxaSE system compared to either of the QIIME based methods. The first principle coordinate axis, PC1 explained 53.22% of variance, the lowest between all three methods.

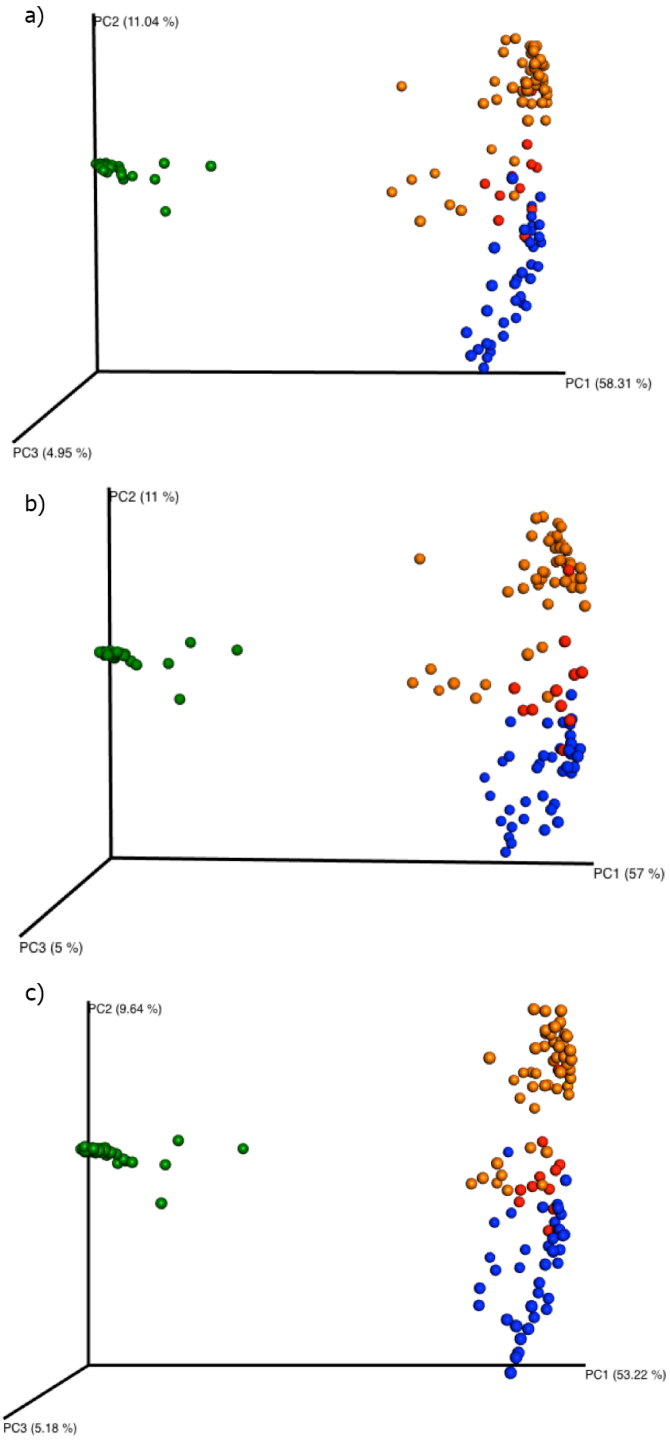


Figure 3–7: Beta diversity principle coordinate analysis plots for distinct taxonomic annotation comparison of sugarcane dataset with a) QIIME at 97% OTU similarity, b) QIIME at 99% OTU similarity and c) TaxaSE. Rhizosphere samples are shown in red, root in blue, soil in orange and stem in green.

3.3.2.3 ADONIS and ANOSIM

ADONIS results for the three methods as listed in Table 3-6 show a slightly different pattern, where the grouping of samples on the basis of environment was best explained by QIIME at 97% OTU similarity with a R^2 value of 0.6797, followed by QIIME at 99% OTU similarity with a R^2 value of 0.671 and lastly TaxaSE, with a R^2 value of 0.622. Overall, the ADONIS results were similar between all three methods.

Table 3-6: ADONIS results for distinct taxonomic annotation comparison between TaxaSE, QIIME at 99% OTU similarity and QIIME at 97% OTU similarity

QIIME at 97% OTU similarity						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	25.417	8.4725	99.008	0.67965	0.001
Residuals	140	11.980	0.0856		0.32035	
Total	143	37.398			1.00000	
QIIME at 99% OTU similarity						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	25.317	8.4391	95.371	0.67145	0.001
Residuals	140	12.388	0.0885		0.32855	
Total	143	37.706			1.00000	
TaxaSE						

	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	23.700	7.9000	76.743	0.62186	0.001
Residuals	140	14.412	0.1029		0.37814	
Total	143	38.112			1.00000	

The ANOSIM results illustrated that for all of the methods, the grouping of samples by environments is statistically significant, with a p-value of 0.001 and R-value closer to +1 (Table 3-7). All three methods generated an R-value of more than 0.8, however TaxaSE produced a slightly lower, but still strong ANOSIM result compared to the other two methods.

Table 3-7: ANOSIM results for distinct taxonomic annotations comparison between TaxaSE, QIIME at 99% OTU similarity and QIIME at 97% OTU similarity

ANOSIM		
Approach	p-value	R-value
QIIME at 97%	0.001	0.8528
QIIME at 99%	0.001	0.8558
TaxaSE	0.001	0.8238

3.4 Discussion

TaxaSE represents an advancement in taxonomic annotation compared to current approaches, with the utilization of a potentially more evolutionary correct sequence similarity measure and its application in a microbial taxonomic annotation pipeline.

Given that the true number of species is unknown for a real dataset, a comparison cannot be made solely on the basis of number of species identified. Nonetheless, the system showed comparable performance in an OTU based analysis, while a higher number of annotations were generated when an OTU independent, per sequence annotation is performed. Given that TaxaSE produced similar patterns with respect to alpha diversity results, the new pipeline is as applicable as other pipelines in assessing alpha diversity in ecological studies.

The microbial community was observed to be more diverse in the case of soil, rhizosphere and root habitats, which are expected to have a high degree of diversity (Kirk et al., 2004; Pinton, Varanini, & Nannipieri, 2001). However, samples from the stem environment were far less diverse. This was primarily due to different species inhabiting plant stems, such as members of the Pantoea genera, which may include endophytic microbes that are beneficial to the growth (Gouda, Das, Sen, Shin, & Patra, 2016) and health of the plant (Miguel et al., 2016) as well as pathogenic bacteria, however a single plant species may play as a host for only a limited number of microbes (Imam, Singh, & Shukla, 2016).

Furthermore, the niche endophyte population is dependent on various factors such as host species and environmental conditions (Gouda et al., 2016).

3.4.1 OTU Comparison

TaxaSE showed comparable performance to the RDP classifier (Cole et al., 2005) within QIIME (Caporaso et al., 2010). For rhizosphere and stem samples, the new pipeline annotated more OTUs compared to QIIME, which provided slightly more annotations for root and soil samples. Shannon diversity index were almost identical for all four habitats across both pipelines.

The beta diversity plots were almost identical, capturing the same ecological patterns. A limited separation is observed between soil, rhizosphere and root samples, with some overlap because of the plant-soil close association (Ke & Miki, 2015). Distinct differences in microbial communities between rhizosphere and bulk soil, as well as root communities that are influenced by plant types and the variety of compounds released by plant roots (Garbeva, Veen, & Elsas, 2004) may have been the source of limited separation between these samples as well.

For ADONIS statistical test, a p-value of 0.001 indicates that grouping of samples by environments is statistically significant at an alpha of 0.05. The R^2 value for TaxaSE was 0.37767, which was slightly higher than R^2 value for QIIME. This

would suggest that the new system was slightly better at explaining variance. ANOSIM results were similar for both approaches.

Overall, TaxaSE provided similar alpha diversity and beta diversity result compared to QIIME in OTU based comparison and illustrates that TaxaSE is a useful alternative to percentage identity based annotation approaches currently employed in taxonomic annotation.

3.4.2 Distinct Taxonomic Annotations

For all four datasets, there is an increase in the number of observed species as well as Shannon diversity index for TaxaSE system, albeit at a slightly higher standard error. While the true number of species is unknown, TaxaSE generated a higher number of taxonomic annotations compared to QIIME based approaches across all four habitats. This illustrated the applicability of an OTU-independent approach as being an alternative method to industry standard OTU based methods. Furthermore, TaxaSE is again capturing similar alpha diversity patterns and the results generated are similar to QIIME's results.

As for beta diversity analysis, QIIME at 97% OTU similarity, QIIME at 99% OTU similarity and TaxaSE, displayed similar patterns and were able to differentiate between different habitats. Furthermore, similar to OTU comparison, here as well stem samples were distinctly separated from root, soil and rhizosphere for all

three methods. Thus, TaxaSE is well suited to identifying ecologically distinct microbial assemblages.

In the case of TaxaSE, slightly less variability was accounted by the first axis, PC1 compared to QIIME at 97% OTU similarity and 99% OTU similarity. This may be because more common taxa were observed for TaxaSE system and therefore the ability of the system to explain variability on the basis of taxonomy fell as an increase in the number of variables leads to a reduction in the total variation explained (Nagelkerke, 1991). A similar case was observed between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity as the latter's first axis explained slightly less variability at 57%, compared to former's 58.31%.

For ADONIS results, QIIME at 97% OTU similarity explained the most variance, followed closely by QIIME at 99% OTU similarity, with TaxaSE explaining the least. The results correlate inversely with the number of distinct taxonomic annotations, where QIIME at 97% OTU similarity produced the least number of distinct annotations and explained the most variance and TaxaSE system produced the most number of distinct annotations but with low explanation of variance. Therefore, given that the ADONIS test described how much variation is explained by grouping on the basis of location, less variation is being explained by approaches with a higher number of taxonomic annotations. This may be because some taxonomic annotations were common across different habitats and approaches such as QIIME at 99% and TaxaSE were able to extract these annotations more in comparison to QIIME at 97%.

The system has certain limitations that need to be kept in view. Considering that TaxaSE uses only the best alignment as part of analysis, a least common ancestor approach may be more suitable for enhanced performance. The best alignment approach may also limit higher resolution, as sequences that are very similar to each other may get annotated in a single taxon, and therefore multiple alignments are needed to elucidate the difference between these sequences. Using SCFG based tools could improve the system performance as these tools are more equipped to handle distantly related taxon compared to similarity based sequence aligners.

3.5 Conclusion

TaxaSE represents a novel approach to taxonomic annotation of microbial DNA. By exploiting evolutionary conservation present in the 16S rRNA gene as well as directly analyzing sequences, it improves upon current methods that rely on percentage identity methods while using an OTU based approach.

The OTU independent approach provides an alternative method to improving taxonomic annotation. While this comes at the expense of more computational time and requirement of higher resources, it can be used to delve deeply into finer level of taxa levels and can lead to improved annotation process as a result. Alpha diversity results also illustrate a similar picture where TaxaSE generated the highest number of annotations across all habitats in comparison to QIIME based methods. This highlights the viability of the new approach.

As computational resources are getting cheaper and more readily available, for finer resolution, this approach can be applied for ecological projects where samples are smaller in size. The results of applied environmental dataset analysis demonstrate the application of using TaxaSE as an alternative to industry standard pipelines such as QIIME, with respect to OTU based comparison, while demonstrating comparable performance in distinct taxonomic annotation based approach. With the ability to annotate individual sequences using a novel scoring approach based on Shannon entropy, TaxaSE represents a step forward in taxonomic annotation of microbial DNA sequences.

Furthermore, by integrating in QIIME, TaxaSE can be used quickly in microbial ecology projects to enhance the resolution of annotation and explore the diversity within each OTU. In essence this would require minimum effort on learning the pipeline and generating ecologically important results. Future work for this pipeline would include more extensive benchmarking across more habitats as well as different OTU algorithms present within QIIME and other taxonomic annotation tools particularly those incorporating entropy such as oligotyping (Eren et al., 2013).

Chapter 4: Assigning Environmental Terms to Sequences using SEQenv

4.1 Introduction

Microbial communities play an essential role in the inner working of every environment on the planet. These microorganisms are genetically diverse and occupy every known habitat where they participate in driving nutrient cycles and form the basis of food webs. The niche of these organisms however, is influenced by environmental characteristics especially in the context of the Baas-Becking hypothesis (Baas-Becking, 1934), which states that, "everything is everywhere but the environment selects (De Wit & Bouvier, 2006)." This determines relative abundance and patterns in diversity of microbial communities.

The advent of next-generation sequencing technology has ushered in a new era of ecological analysis. There is an increasing interest in comprehensive description of environmental context and experimental methods used for sequencing data, In the absence of this, such data sets would be of less value for comparative studies or discovering linkages between genetic potential and the diversity and abundances of organisms (Field et al., 2008). Furthermore, a full understanding of the role of environmental selection of microbial diversity can only be realised if associated metadata related to geographical or environmental information can be exploited. Given that microbes affect the inner working of the environment directly via participating in various functional processes, environmental data

concerning these sequences would be more informative when understanding the various factors that influence their diversity (Lombardot et al., 2006). This can grant the ability to extract patterns that may not be visible when abundance and diversity data is viewed in isolation and without context.

To that end, various formal specifications and guidelines have been developed to facilitate curation of metadata in a standardised format such as the minimum information about any sequence specifications (Yilmaz et al., 2011) by the Genomic Standards Consortium (Field et al., 2011). Furthermore, sequence data submission to many public databases including GenBank (Benson et al., 2012) and INSDC (Nakamura, Cochrane, & Karsch-Mizrachi, 2013) as well as online bioinformatics tools like MG-RAST (Aziz et al., 2008) have specific metadata fields for storing contextual information concerning the sequences. Moreover, large scale projects such as the Earth Microbiome Project (Gilbert et al., 2014), which aim to develop a global catalogue of microbial diversity, store contextual metadata information as well.

As a consequence of metadata acquisition and availability, newer approaches have been developed that utilise this information in a novel way. The foremost is microbial biogeography, which emerged to link microbial diversity with geographical locations and aimed to determine the various distribution patterns of microbes (Martiny et al., 2006). However, while most work in biogeography has been habitat-specific, environmental factors rather than geographical locations may be more influential on microbial diversity (Fierer & Jackson, 2006) and represents the current limitation in these approaches. Other examples of

metadata use include visualization of phylogenetic trees with environmental context (Pirrung et al., 2011) and linking publicly accessible metadata to sequencing reads (Nayfach & Pollard, 2016; Sunagawa et al., 2015). Hence, the addition of environmental annotation, which serves as a descriptor of the habitat, to taxonomic identities can be a significant improvement to analysis capability and provide a more in-depth approach towards understanding microbial communities rather than application of geographical location data, as is the case with microbial biogeography (Fierer & Jackson, 2006).

Before applicable environmental annotation can be performed for sequences, a precise and consistent environmental description for the origins of these sequences and the samples they came from, is needed. To that end, the Environmental Ontology, or ENVO Ontology provides a structured, controlled vocabulary in a hierarchical list of descriptors, which can then be used to organize environmental data in a coherent and unambiguous manner (P. L. Buttigieg et al., 2013). In essence, the ontology provides a list of standardized environment descriptors that can be used to properly explain the environment or habitat as well as its noticeable features and has been adopted by MG-RAST (Aziz et al., 2008), the iMicrobe project (Pier Luigi Buttigieg et al., 2016) and Earth Microbiome Project (Gilbert et al., 2014).

The NCBI-NT database provides a wealth of information with respect to environmental metadata. Sequences submitted to the database may contain a GenBank (Benson et al., 2012) metadata field known as *isolation source*, which provides the environment source from where the organism DNA was extracted

from ("The GenBank Submissions Handbook [Internet]," 2011-). This can then be exploited to label sequences with the necessary environmental annotation and can enable characterization of any ecological project with respect to environmental terms using the ENVO ontology.

SEQenv (Sinclair; Sinclair et al., 2016; Sinclair L, 2016) is a new, cutting edge pipeline, which can generate environmental information for sequences, primarily using the isolation sources metadata field from NCBI-NT. The pipeline begins by retrieving highly similar sequences from the NCBI-NT database using the BLASTN algorithm (Altschul et al., 1990). From the hits that match against the query sequences, text fields carrying environmental information such as isolation sources found in the metadata are extracted. Given that isolation sources are in the form of short English sentences, this information is converted into the nearest ENVO ontology terms (P. L. Buttigieg et al., 2013). Text mining is therefore performed on the extracted information, which identifies ENVO terms such as “glacier”, “soil” or “forest”. The pipeline is uniquely placed to derive environmental annotations for sequences as so far, no automated bioinformatics pipeline exists for this purpose. Lastly, the pipeline can be used for both nucleotides and protein sequences (Sinclair et al., 2016). However, SEQenv is only able to generate a list of environmental terms on the basis of sample datasets and lacks a taxa centric approach to environmental annotations. Hence, environmental annotations at sequence level are not provided. Such information is critical to identify niches for particular taxa and their potential role in driving ecosystem functions.

This chapter aims to address these deficiencies by integrating SEQenv with the TaxaSE system developed in Chapter 3 and extending the pipeline itself by way of a taxa centric approach, which would be valuable to any biologist seeking to understand the various taxa present in the habitat and which environment they originated from, enabling a more thorough analysis of which taxa are abundant in certain habitats and recovery of patterns in taxon distribution across different habitats and environmental gradients.

The extension consists of two parts, each providing environmental annotations under a different context, with first part providing taxa abundance on a per term basis while the second part lists environmental term abundance under a per taxa context. This chapter therefore tested the hypothesis that environmental annotation could enhance analysis of microbial communities and the annotations generated were in accordance with prior knowledge in the literature about the habitats the microbes belong to.

Two real amplicon datasets belonging to distinct biomes were selected in order to determine the applicability of both the SEQenv pipeline and the newly developed extension, towards environmental annotation of datasets belonging to different habitats. Lastly results were visually illustrated for improved readability. This was accomplished by illustrating the abundance of environmental terms or taxa in larger fonts, which was more effective at revealing the most essential and interesting information in quick fashion and therefore identity trends or patterns more clearly as compared to listing data in a tabular format.

4.2 Materials and Methods

4.2.1 SEQenv Pipeline

The SEQenv pipeline was utilized to generate environmental annotations for the datasets by way of developing a global matrix of sequences and environmental terms. The process followed by SEQenv is given as below:

1) For every OTU in a list of query OTUs, BLASTN was used to search against the NCBI-NT database. By default, the top 10 hits for each OTU are stored.
2) From the hits, isolation source metadata was extracted using the NCBI global identifier.
    a. Genomes, genes and sequences submitted to NCBI have associated metadata available. Isolation source describe the geographical and/or environmental information related to the specific sequence or genome that was submitted to NCBI.
    b. However, given that isolation sources are not available for every sequence, either because no information was submitted or is available, global identifiers that do not have this information were ignored.
3) Isolation sources are small text fields containing short sentences that describe the environment from where the organism was isolated from. The ENVO tagger portion of the SEQenv pipeline was run on this metadata to extract equivalent ENVO ontology terms.

4) A query OTU may have one or more isolation sources depending on the number of hits against NCBI-NT. The resultant ENVO terms were normalized for every query sequence. In essence, each sequence was described by a set of ENVO terms and their associated frequency, and for the whole dataset, a matrix of sequence-term was created. For each sequence, the ENVO terms were normalised by assigning weights to them. By default, the "flat" approach was used, where weights were calculated according to raw occurrence counts. This ensured that no environmental information was discarded in this process.
5) Finally, at sample level, the terms were aggregated but weighted by OTU abundances in the sample.

An example of this is illustrated in Figure 4-1.

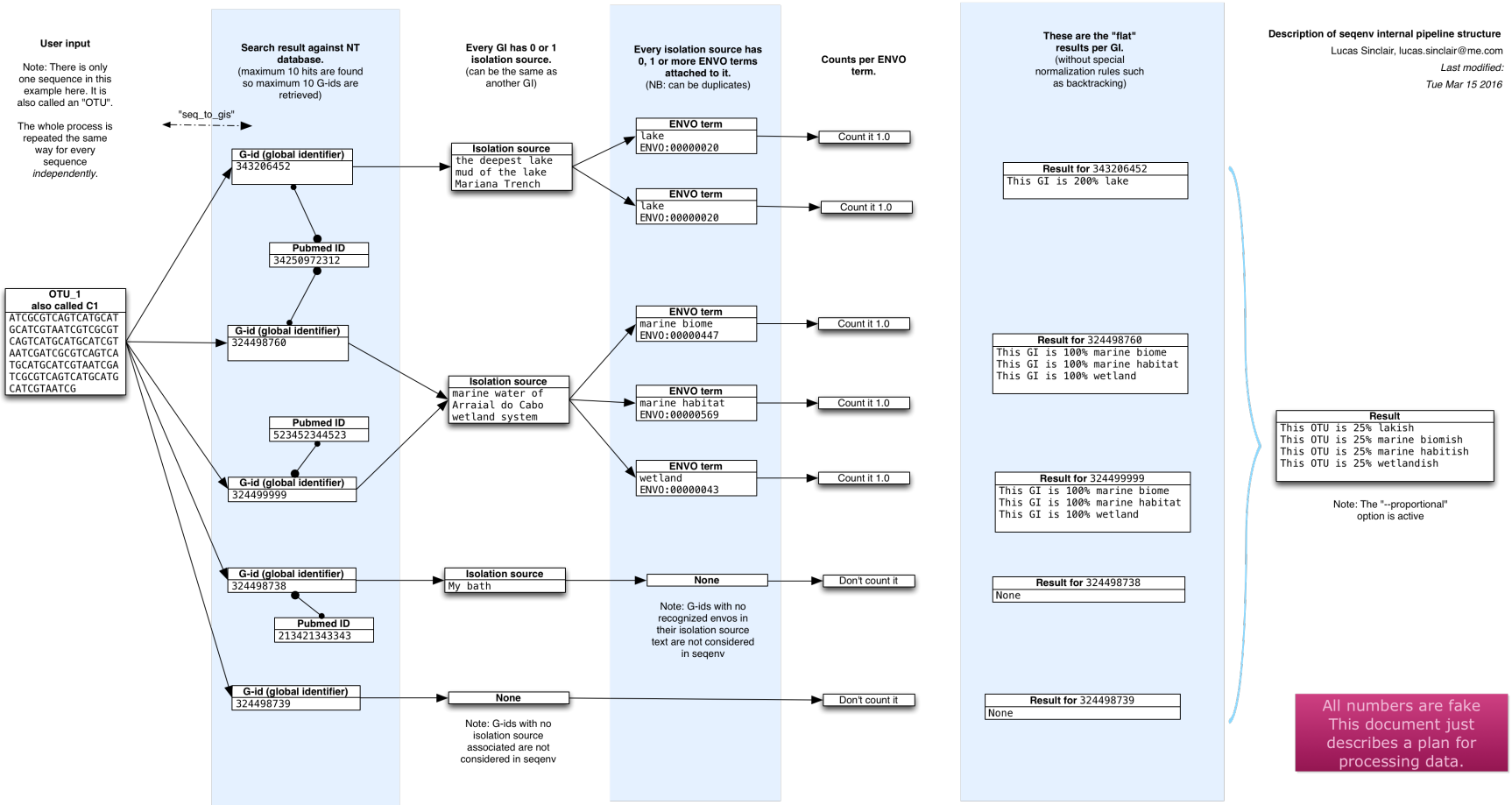


Figure 4–1: SEQenv process diagram illustrating the various steps taken to generate ENVO terms for an OTU. Modified from (Sinclair)

4.2.2 Integration with TaxaSE

Given that TaxaSE followed an OTU independent approach, new tools (as listed in the appendix) were developed here to select unique sequences from taxonomic annotation results that can then be given to SEQenv pipeline for environmental tagging. The approach followed is given as below:

1) From the list of distinct taxonomic annotation results, a collection of sequences was selected on the basis of a genus level threshold.
2) Relative abundances of the taxa were generated for every annotation result.
3) Sequences belonging to every taxa were randomly selected. The number of sequences selected was directly proportional to the relative abundance in the collection of the sequences. For example, a genus with higher relative abundance had more sequences selected from it compared to a genus with lower relative abundance. As SEQenv pipeline uses BLAST, this was done in order to reduce computational resource requirements.
4) The random selection ensured that a wide variety of sequences were used for analysis and were representative of the sample diversity.

The integration of SEQenv pipeline is illustrated in Figure 4-2. This provided a single integrated approach for both taxonomic and environmental annotation of sequences.

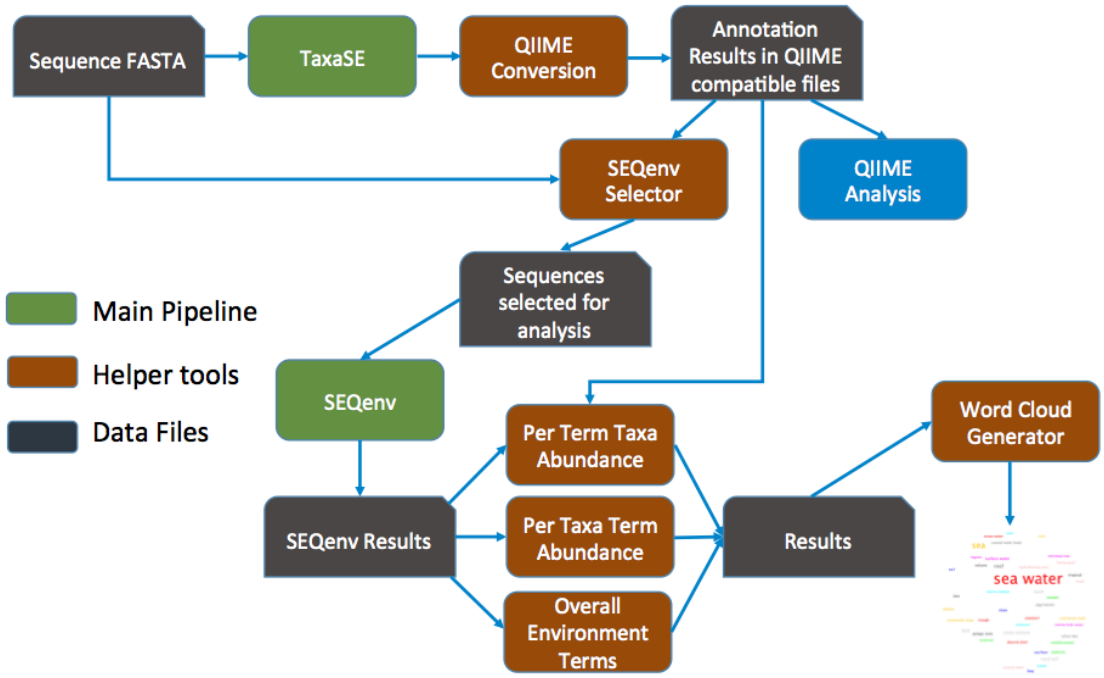


Figure 4–2: Integration and enhancement of SEQenv system, with pipelines shown in green, helper tools in brown and data files in black

4.2.3 Per Environment Term Taxa Abundance

A taxon abundance or contribution to each environment terms provides more detailed information and can help understand which sequences may be more important in contributing to a particular environment. Building upon the current version of SEQenv (Sinclair L, 2016) by extracting taxa abundance for a given environmental term, allows for the opportunity for detailed analysis of the partitioning of diversity across habitats within the context of the samples being analysed.

SEQenv results consist of a global matrix between sequences and their associated environment terms. Given that taxonomic annotation information is present for these sequences via the TaxaSE system, a per environment term taxa abundance result can therefore be generated. In essence, for every environment term, a list of ranked taxa is produced. The ranking of these taxon is dependent upon how much they contributed to the specific term.

1) For every environment term, select the sequences that contribute towards it. These were taken from SEQenv results.
2) For each of these sequences, the associated taxonomic annotation information was recovered from TaxaSE results.
3) If one or more sequences belong to the same taxonomic annotation, the contribution by each sequence was added together.
4) Taxonomic annotations were then ranked according to how much they contributed to the environment term.

For the most abundant environment terms, word clouds were then generated. These word clouds represent the taxonomic annotation in the descending order of magnitude with respect to how much each annotation contributed to the term. It is important to note here that a sequence can have multiple isolation sources, and therefore can come from a variety of environments. While there may be repeats of taxonomic annotations, the magnitude is dependent on the contribution factor for each term and thus can be used to determine how much individual taxa contribute to the signature of specific environments.

4.2.3 Per Taxa Environment Term Abundance

Relating environmental information to sequences in a direct fashion would improve our understanding of how they are distributed across various environments and would be a valuable asset for any biologist aiming to understand the natural habitats and niche specificity of these microbes. Hence, using the same global matrix acquired from SEQenv, a list of sequences and the environments they belong to was created in the following manner:

1) A list of environmental terms for every sequence was generated from SEQenv results.
2) Taxonomic annotation information from TaxaSE system results was recovered and the sequences were assigned the corresponding taxonomy.
3) If one or more sequences had the same taxonomy, the contribution by each environmental term was added together.
4) Environmental terms were then ranked according to how much they contributed to the taxa.

Results were stored in a text file, which listed the sequences by abundance in descending order. For selected four taxa, pie charts were then used to illustrate the various environments in which taxa may exist. Similar to per term taxa abundance, a sequence can have multiple isolation sources and therefore can come from multiple environments, the ranking of which depends on how much each environmental term contributes to the sequence.

4.2.4 Datasets

In order to illustrate the effectiveness of SEQenv system when combined with TaxaSE in determining environmental information related to sequences and to show its applicability, datasets belonging to distinct and diverse biomes were selected. These datasets included soil, rhizosphere and plant microbiome from sugarcane (*Saccharum* spp.) sequenced by Dr. Kelly Hamonts at HIE, Western Sydney University, and samples from two distinct marine sub habitats (Jeffries et al., 2015). These were the same datasets as used in Chapter 3 of this study. The number of sequences selected from these datasets is given in Table 4-1.

Table 4-1: Datasets selected for analysis with enhanced SEQenv system

Habitat	Sub habitat	Total number of sequences
Sugarcane	Rhizosphere	3000
	Soil	3000
	Stem	3000
	Root	3000
Marine	Coral Atoll	1500
	Southern Ocean	1500

SEQenv version 1.1.0 was run with default parameters with these sequences using BLASTN. The parameters and the values used are listed in Table 4-2.

Table 4-2: Parameters used for SEQenv analysis.

SEQenv Parameters		
Parameter	Information	Default value used
--min_identity	Minimum identity in similarity search.	0.97
--min_coverage	Minimum query coverage in similarity search.	0.97
--proportional	Should we divide the counts of every input sequence by the number of envo terms that were associated to it.	True
--search_db	The path to the database to search against.	nt
--max_targets	Maximum number of reference matches in similarity search.	10
--seq_type	Either `nucl` or `prot`.	nucl (nucleotide)
--search_algo	Either 'blast' or 'usearch'.	blast
--e_value	Minimum e-value in similarity search.	0.0001

In the context of this study, analysis of the datasets was divided into three sections:

- *Per Habitat Environmental Terms*: This represents the environmental terms as generated by the main SEQenv pipeline.
- *Per Environmental Term Taxa Abundance*: This represent the taxonomic abundance as generated by the first part of the new extension to the SEQenv pipeline. Furthermore, SEQenv results for each sub-habitat from

the aforementioned datasets were aggregated and Per Term Taxa Abundance results were then generated from the resultant information.

- *Per Taxa Environmental Term Abundance*: This represents the environmental terms abundance on a per taxa basis, as generated by the second part of the new extension to SEQenv pipeline. Similar to per environmental term taxa abundance, SEQenv results for each sub-habitat were aggregated.

4.2.5 Word Cloud Generation

Given that the SEQenv pipeline is under development, the latest version does not have the ability to generate a word cloud to represent the relative abundance of environmental terms or per term taxa abundance, which necessitated the development of an additional word generation tool to integrate this into TaxaSE.

Environmental terms or taxa were illustrated on the strength of their abundances, with higher abundance producing a larger font size. Colours were randomly selected to improve readability against a white background, while a circular pattern was used for illustration purposes. Additionally, for per term taxa abundance word clouds, instead of full taxonomic annotations, the lowest two taxa levels were used to illustrate the taxon found. Table 4-3 also describes the word cloud java tool.

4.2.6 List of Tools

The extension to the SEQenv pipeline was developed in Java programming languages and was represented by a collection of tools. The list of tools, including word cloud generation and their description is listed in Table 4-3.

Table 4-3: List of tools

Tool	Description
seqenv-selector.jar	This tool selects sequences based on taxonomic annotation and abundance data present in TaxaSE system results.
seqenv-abd.jar	This tool aggregates the results on SEQenv pipeline on a per-term basis and generates a list of most abundant environmental terms.
seqenv-cloud-gen.jar	As the current version of SEQenv does not provide word cloud generation functionality, this tool was developed to illustrate the pipeline results.
seqenv-rev.jar	These tools represent the extension to the SEQenv pipeline. By using the TaxaSE annotation information and SEQenv results, this tool is able to generate both per term taxa

	abundance and per taxa term abundance results.

4.3 Results

4.3.1 Per Habitat Environmental Terms

4.3.1.1 Sugar Cane Dataset

The environmental terms for the sugarcane dataset are illustrated in Figure 4-3. Samples belonging to rhizosphere showed the environmental term “soil” as being the most prevalent (Figure 4-3a). Other similar terms were also observed, such as “rhizosphere”, “forest soil”, “prairie” and “agricultural soil”. Of importance was the occurrence of the environmental terms such as “activated sludge”, “garden” and “contaminated soil” as more taxa with these metadata were prevalent in these datasets. SEQenv was unable to generate environmental terms for the ENVO IDs 1000196, which stood for “coniferous forest biome”, 446, which was “terrestrial biome” and lastly 447, which was “marine biome”. This may be due to limitation in the SEQenv pipeline.

The list of environment terms for soil samples were also similar to rhizosphere samples, with the “soil” term being the most significant environmental term observed (Figure 4-3b). However, “forest soil” was observed relatively strongly here compared to rhizosphere results. Similar to rhizosphere results, SEQenv was unable to convert a few ENVO IDs to their corresponding environmental terms. These included 447, which was “marine biome”, 1000181, which was “mangrove biome”, 428, which was simply “biome” and 2030, which was “aquatic biome”.

For the stem samples, the environmental term “garden” was strongly observed compared to other terms, with exception of “soil” (Figure 4-3c). Other important terms included “forest soil”, “biofilm” and “garden soil”. A few environmental terms were not properly generated. These include 1000196, which stood for “coniferous forest biome”, 1000047, which was “mediterranean sea biome”, 447, which was “marine biome” and finally 1000181, which was “mangrove biome”.

Lastly, the root samples showed similarity to both soil and rhizosphere samples, as the environmental term “soil” was the most observed term in these three sub-habitats (Figure 4-3d). As illustrated in the ANOSIM results for TaxaSE in Chapter 3, where the grouping of samples by environment were statistically significant, these three sub-habitats were close to each in taxonomic makeup and therefore would contain sequences that would have an environmental term of “soil” as the isolation source. Similarly, beta diversity plot for TaxaSE also illustrate samples from these sub-habitats, while grouped together individually, were closer to each other compared to samples from stem sub-habitat. While the plots show similar list of environmental terms, with exception of the few most strongly observed terms, the ranking of the terms themselves vary across these habitats. Undetermined ENVO IDs included 446, which was “terrestrial biome”, 447, which was “marine biome”, 1000196, which was “coniferous forest biome” and 2030, which was “aquatic biome”.

Overall, the environmental term “soil” was prevalent across all sub-habitats, however other terms were ranked differently. Stem sub-habitat was more unique compared to soil, rhizosphere and root.

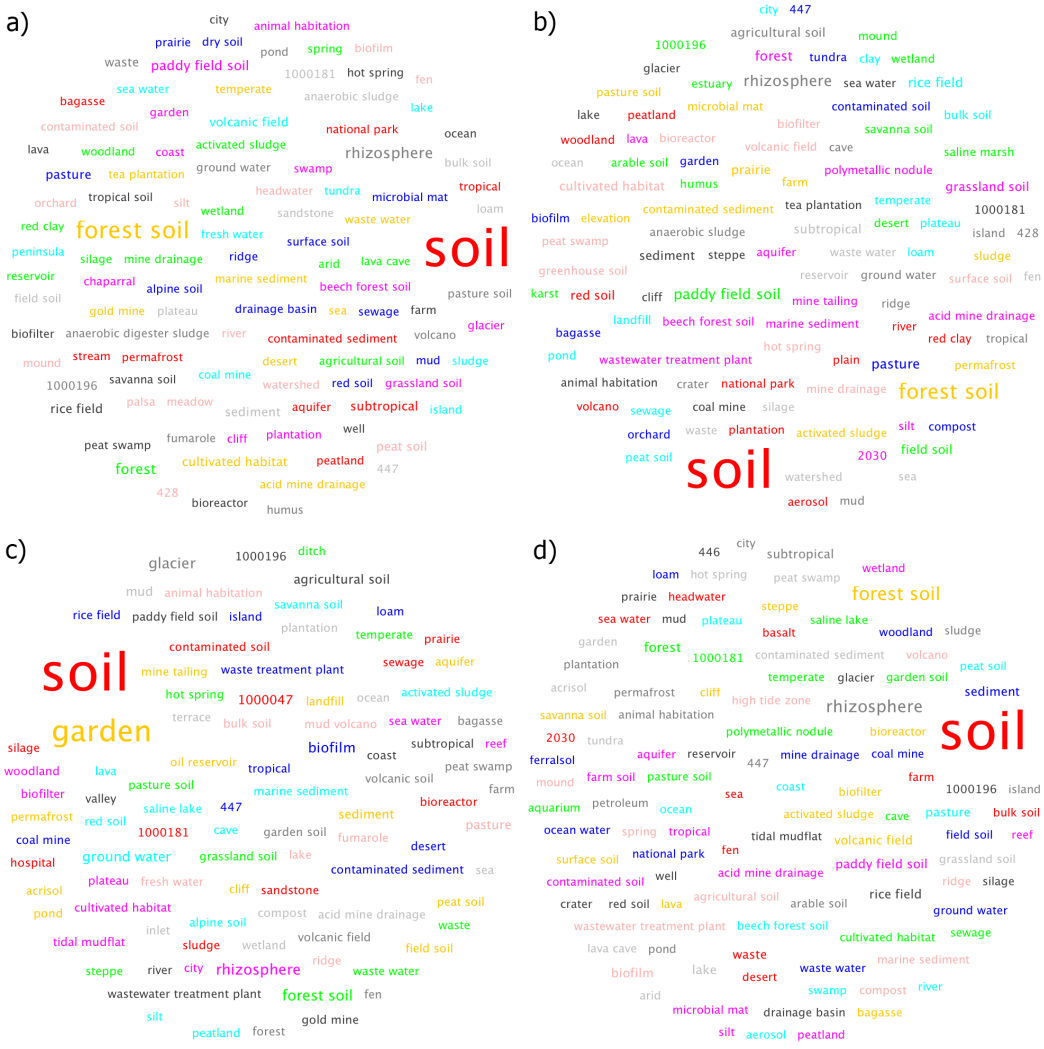


Figure 4–3: Environmental terms generated for the sub-habitats a) rhizosphere b) soil c) stem and d) root. More abundant terms are highlighted with larger font.

The top 10 environmental terms ranked according to their abundances for each sub-habitat is listed in Table 4-4, with the differences between sub-habitats highlighted in bold. The unique terms here are those environmental terms that only exist in the specific sub-habitat.

Table 4-4: Top 10 environmental terms observed in sub-habitats from the sugarcane dataset, sorted in a descending order of abundance and unique terms highlighted in bold.

	Sub-habitats			
Rank	**Soil**	**Rhizosphere**	**Root**	**Stem**
1	soil	soil	soil	soil
2	forest soil	forest soil	forest soil	**garden**
3	rhizosphere	rhizosphere	rhizosphere	**glacier**
4	paddy field soil	forest	forest	**ground water**
5	rice field	paddy field soil	paddy field soil	forest soil
6	forest	pasture	sediment	biofilm
7	pasture	volcanic field	rice field	rhizosphere
8	cultivated habitat	rice field	pasture	pasture
9	sediment	subtropical	biofilm	agricultural soil
10	subtropical	cultivated habitat	cultivated habitat	mud

4.3.1.2 Marine Dataset

For the coral atoll marine samples, the environmental term “sea water” was the most observed term, with “sea” coming up after that (Figure 4-4a). A few other terms of importance include “bay”, “coral reef”, “ocean” and “sediment”. Similar to samples from sugarcane dataset, SEQenv was unable to determine the

environmental term for the IDs 428, which was “biome”, 447, which was “marine biome” and 1000047, which was “mediterranean sea biome”.

Southern ocean samples also showed a similar list of environmental term (Figure 4-4b). Here as well “sea water” and “sea” terms were the most observed, however “brine pool” was relatively strongly observed here compared to samples from coral atoll. The list of ENVO IDs not mapped onto the proper environmental terms included 447, which was “marine biome” and 1000048, which is “ocean biome”.

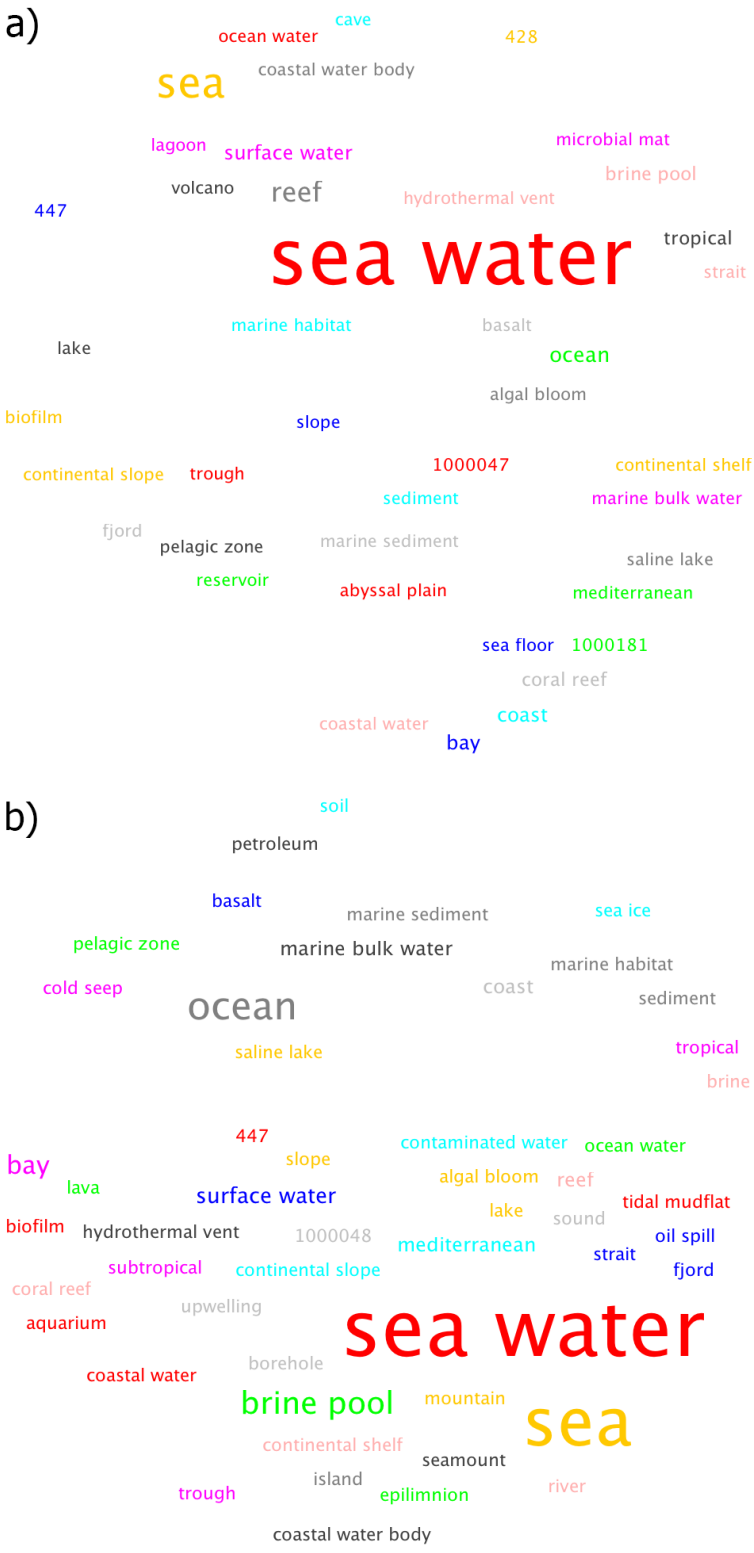


Figure 4–4: Environmental terms for the marine sub-habitats a) Coral Atoll and b) Southern Ocean. More abundant terms are highlighted with larger font.

While both marine samples showed a similar list of environment terms, these differed in the ranking of the terms themselves, which is illustrated in Table 4-5. Here, the ranking of top level environmental terms was the same for both coral atoll and southern ocean samples, however differences were observed in the lower ranked terms where "coral reef" was observed for coral atoll samples while southern ocean had environmental terms like "Mediterranean" and "marine bulk water", which were absent in coral atoll samples.

Table 4-5: Top 10 environmental terms observed in sub-habitats from marine dataset, sorted in a descending order and unique terms highlighted in bold

	Sub-habitats	
Rank	**Coral Atoll**	**Southern Ocean**
1	sea water	sea water
2	sea	sea
3	reef	ocean
4	ocean	brine pool
5	surface water	bay
6	bay	surface water
7	coast	**mediterranean**
8	brine pool	reef
9	tropical	**marine bulk water**
10	**coral reef**	coast

4.3.2 Per Environment Term Taxa Abundance

4.3.2.1 Sugarcane Dataset

While the environment terms "Soil" and "Forest Soil" were similar, the sequences that contribute to these terms differed (Figure 4-5a and 4-5b respectively). This was quite apparent in the differences between both word clouds where the most abundant taxa for "soil" term included *Acidothermus* and *Chloroplast* while *Variibacter* and *Acidobacteriaceae* were more strongly related to the "forest soil" term.

"Rhizosphere" environmental term had *Burkholderia* as being the most abundant taxa while *Acidothermus* was almost non-existent in this case as shown in Figure 4-5c. *Burkholderia* was followed by *Catenulispora sp. Neo1*, *Acidobacteriaceae (Subgroup 1)* and *Dyella*. *Xanthomonadaceae* and *Catennulispora* were also observed, though at a lower abundance.

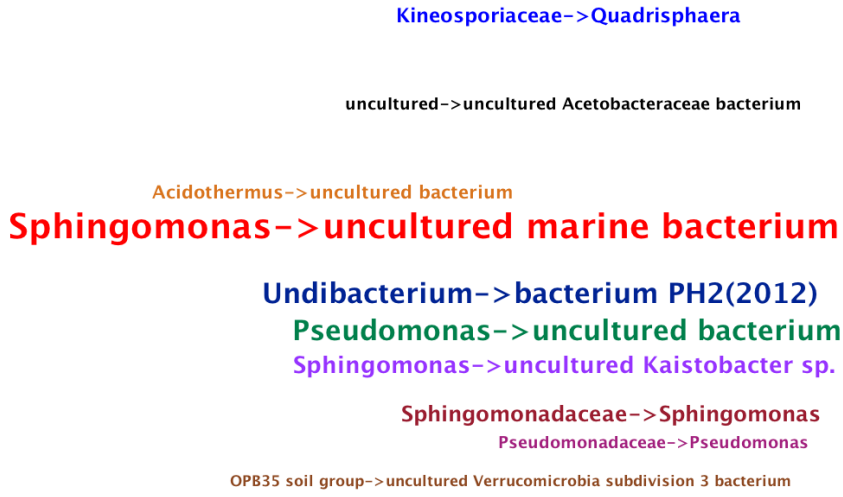
The "garden" environmental term had distinct taxa, which were not observed in other environmental terms (Figure 4-5d). Members of *Pantoea* genus were strongly observed here, while being absent in other environmental terms.

Figure 4–5: Per Term Taxa Abundance for the environmental terms a) soil b) forest soil c) rhizosphere and d) garden. More abundant taxa are highlighted with larger font.

"Contaminated soil" term is an example of significantly different collection of taxa (Figure 4-6a). While not listed in the top 10 environmental terms for the sugarcane dataset, it and "waste" environmental term consists of important collection of taxa that may be relevant to biologists studying these specific taxa. Here, *Sphingomonas*, *Pseudomonas* and *Undibacterium* were more abundant, in that order.

For the “waste” environmental term, *Acidothermus* was the most observed taxa, followed by *Acidobacteriaceae (Subgroup 1)* (Figure 4-6b). Additionally, members of *Chitinophagaceae* family were also seen under this environmental term.

a)



b)

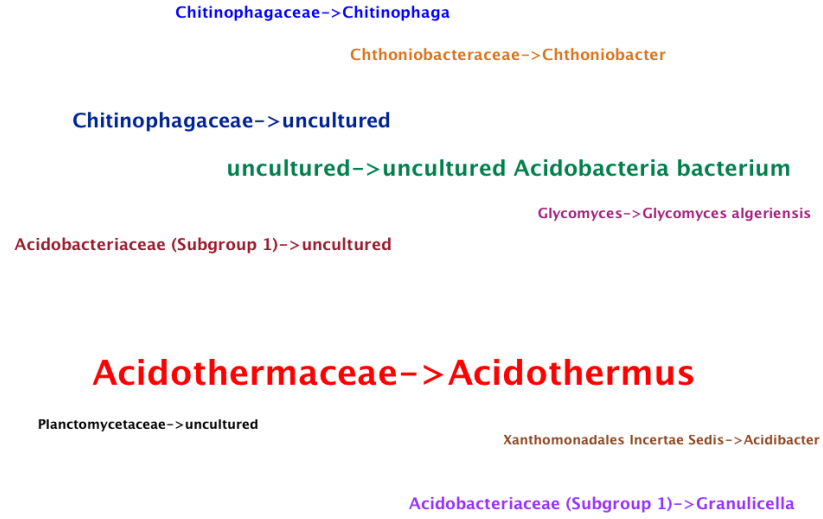


Figure 4-6: Per Term Taxa Abundance for environmental terms a) contaminated soil and b) waste. More abundant taxa are highlighted with larger font.

4.3.2.2 Marine Dataset

Prochlorococcus dominated the taxa abundance for the environmental term "sea water" (Figure 4-7a). The other taxa such as *SAR11 clade* and *SAR86 clade* were also observed, though at lower abundances. On the other hand, while similar taxa were observed for the environmental term "sea", the relative abundance of these taxa were significantly different (Figure 4-7b). *SAR11 clade* and *Synechococcus* became more abundant, while *Prochlorococcus* was observed to be far less prominent than what was observed for the environmental term "sea water".

"Ocean" environmental term showed *Chloroplast* becoming more abundant compared to per term taxa abundance for "seawater" and "sea" environmental terms (Figure 4-7c). Similar behaviour was seen for *Marinimicrobia (SAR406 clade)* as well, which was very low in abundance in the aforementioned environmental terms.

SAR86 Clade and *Prochlorococcus* jointly dominated the "Brine Pool" environment term (Figure 4-7d). Furthermore, a few taxa such as *SAR324 clade (Marine group B)* and *Alteromonas* were also observed, although at a very low abundance. Species diversity was observed to be quite low in this case as only a few taxa contributed to this environment term. Overall, most of the taxa belonged to *Proteobacteria* and *Cyanobacteria* phyla.

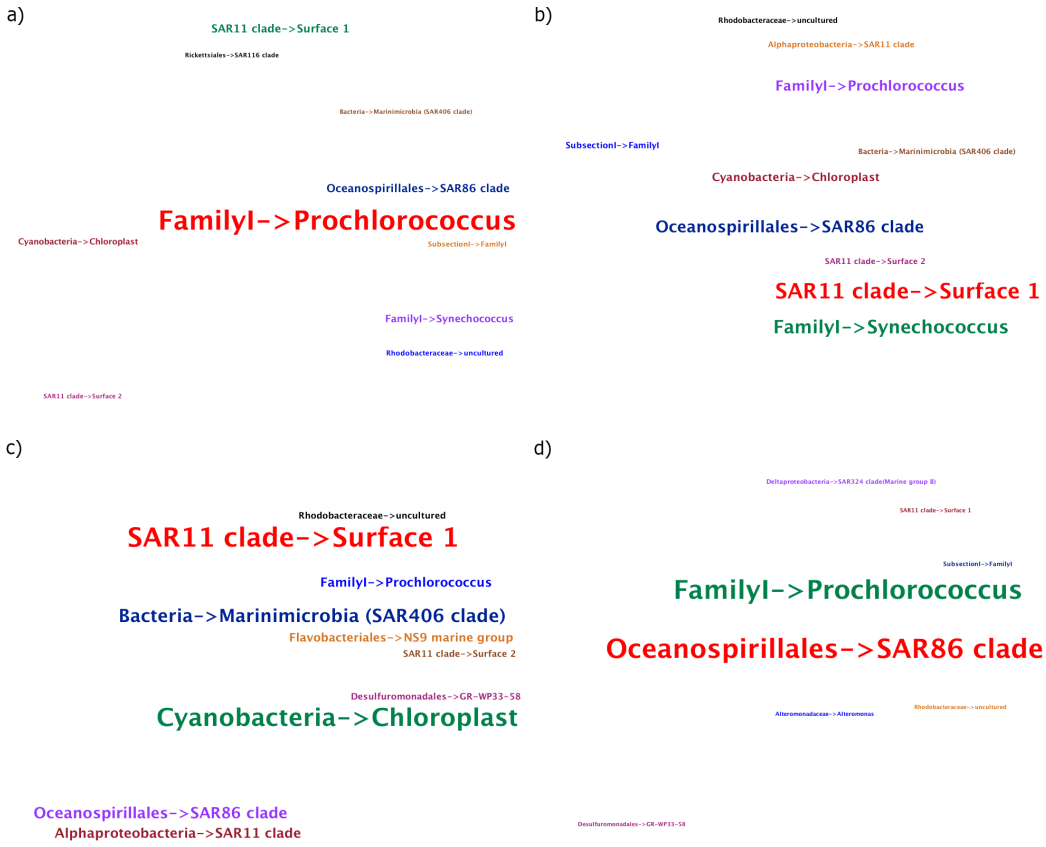


Figure 4–7: Per Term Taxa Abundance for the environmental terms a) sea water b) sea c) ocean and d) brine pool. More abundant taxa are highlighted with larger font.

4.3.3 Per Taxa Environmental Term Abundance

4.3.3.1 Sugarcane Dataset

Per taxa environmental term relative abundance for *Acidothermus* and *Burkholderia* are illustrated in Figure 4-8. While the environmental term “soil” dominated the list of terms for both genera, 52.6% for *Acidothermus* and 45.7% for *Burkholderia*, differences were observed for the lower ranked terms.

"Forest soil" was the second most observed term for *Acidothermus* at 9.9% (Figure 4-8a), however for *Bulkholderia* the term "rhizosphere" was observed higher than "forest soil", accounting for a significant portion of the environment terms at 20% and "forest soil" term accounting for 10% here, similar to *Acidothermus*.

The terms "woodland", "waste" and "rice field" were ranked higher for *Acidothermus* (Figure 4-8a) as well, at 5.5%, 4.9% and 4.8% respectively. For *Bulkholderia*, the "waste" term was not in the top 6 environmental terms (Figure 4-8b), and furthermore the terms "woodland" and "rice field" were not observed at all for this genus.

"Field soil", "peat swamp" and "sludge" environmental terms were observed for *Bulkholderia* at 3.1%, 2.6% and 2.1% respectively, however they were absent from the collection of top 6 terms for *Acidothermus*. Lastly, the remaining collection of environmental terms came at 15.5% for *Acidothermus* and 16.4% for *Bulkholderia*.

Overall, distinct differences were observed between both genera. For *Acidothermus*, with exception of the most abundant "soil" term, others gradually decreased in how much they accounted for in the list of environmental terms. However, *Bulkholderia* showed the gradual decrease after the third ranked "forest soil" term.

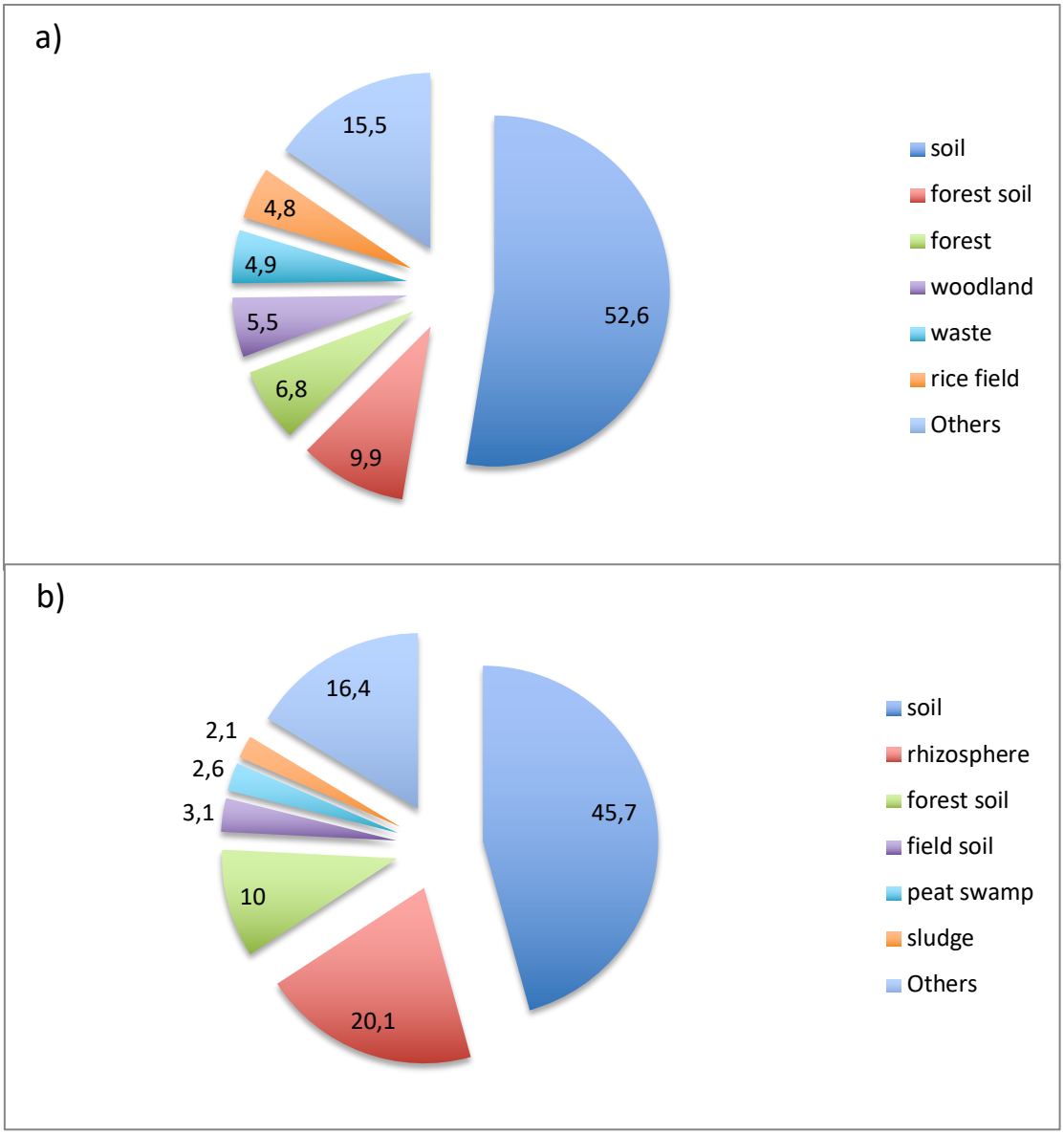


Figure 4–8: Per Taxa Term Abundance for a) Acidothermus and b) Bulkholderia. Top 6 environmental terms are illustrated with the pie chart.

4.3.3.2 Marine Dataset

The per taxa term abundance pie charts for the genus *Prochlorococcus* and *Synechococcus* are illustrated in Figure 4-9. For *Prochlorococcus*, the environmental terms "sea water" was the most observed term, accounting for

73.35% of environmental terms, an overall majority (Figure 4-9a), which came down to third rank for *Synechococcus*, at 26.5%.

Furthermore, "reef" was observed strongly for *Synechococcus* at 27%, however the term was absent in the top 4 list for *Prochlorococcus*. Furthermore, "ocean" term was present for *Prochlorococcus* at 1.49%. Other differences included the term "brine pool" at 9.17% for *Prochlorococcus*, although it was absent for *Synechococcus*. Lastly, "coast" environmental term was observed only for *Synechococcus* at 2.7%.

Overall, environmental term distribution was different between both genera. A single "Sea water" term dominated *Prochlorococcus* list of environmental terms while *Synechococcus* saw three terms accounting for most of the environmental terms observed, on an almost equal level and where "reef" term was distinctly observed for *Synechococcus*.

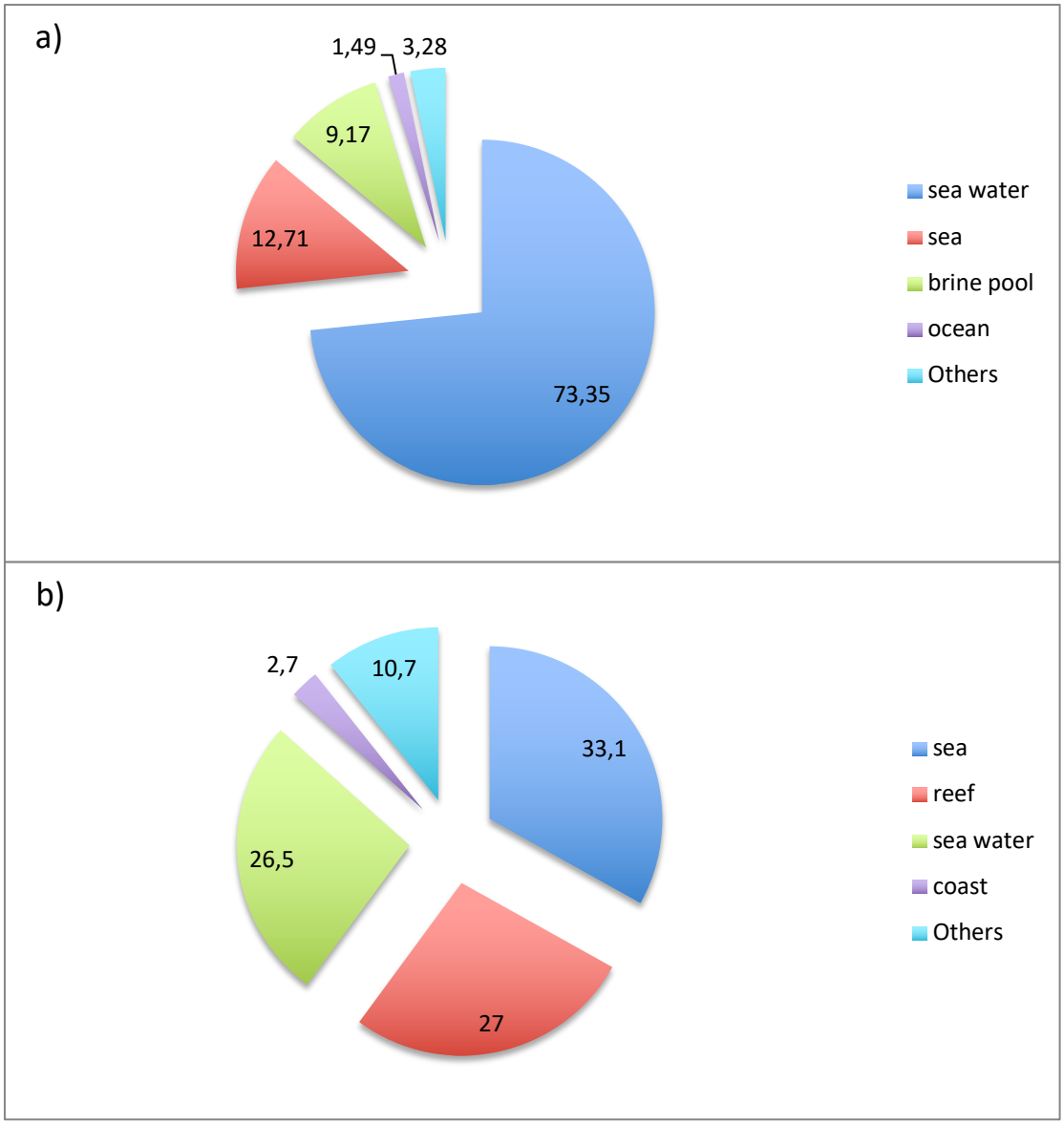


Figure 4–9: Per Taxa Term Abundance for a) Prochlorococcus and b) Synechococcus. Top 4 environmental terms are illustrated with the pie chart.

4.4 Discussion

Sequence annotation can now be enhanced with environmental data, by way of exploiting information available in associated metadata in databases such as NCBI-NT. This can in turn provide a more in-depth view into the microbial community and a more effective approach towards analysis for many ecological projects.

The analysis of the various habitats illustrates the effectiveness of the new extension to SEQenv. Significant patterns emerge where distinct taxa were strongly observed on the basis of the environment origin. By effectively linking observed taxa to environmental terms, the system produces an ecologically important perspective into the analysis of 16S rRNA gene sequences and enables a more thorough approach to environmental annotation of sequences, aiding in interpretation of taxonomic annotation.

Cases were found where the SEQenv pipeline (Sinclair L, 2016) was unable to resolve the environmental term at a deeper level, such as for the environmental terms “soil” and “sea”. Given that “soil” term exists at a higher level than other terms such as “forest soil” in ENVO ontology (P. L. Buttigieg et al., 2013), it is more likely that the isolation sources for these sequences were not detailed enough to determine the precise environment they were isolated from. SEQenv selected a higher level of environment term instead, as the metadata could not provide more specific details about the environment. Additionally, some ENVO IDs could not be

resolved to the proper environmental terms, which may be a limitation of the tool as some of these IDs were similar across various sub-habitats.

4.4.1 Per Habitat Environmental Terms

4.4.1.1 Sugarcane Dataset

While the results for root sub-habitat were similar to soil and rhizosphere, differences were observed for the presence of environmental terms such as "sediment" and "biofilm", which were ranked higher. This might be because of the taxa that belonged to these terms was more abundant in the root habitat due to plant-soil close association (Garbeva et al., 2004). Furthermore, biofilms play an important role in plant-microbial interactions in the rhizosphere (Danhorn & Fuqua, 2007). Additionally, "forest soil" environmental term was relatively more prominent for soil samples compared to rhizosphere results and this might be due to difference in abundance of taxa that are more prevalent in forest soils, which are located further away from the phytobiome system (Garbeva et al., 2004).

The differences between stem samples and others, which was driven by terms such as "garden", "glacier water" and "ground water" can be explained by these terms being driven by taxa unique to the stem habitat and likely to be endophyte in nature. These taxa live within the plant biomass in a symbiotic relationship (Gouda et al., 2016) and therefore observed in the samples taken from the stem. This may be a result of comparatively lower number of sequences from these habitats exist in the database. Nonetheless, given that SEQenv (Sinclair L, 2016)

acquires isolation sources based on the sequences in the dataset, the differences in species found in the stem samples compared to other samples led to strong difference in environmental based tagging. Furthermore, as the stem samples were taken from the stem of sugarcane plants, the ranking of environmental terms in this case are a good representative of the type of the environment the microbial sequences came from. This highlights the value of SEQenv in discriminating between habitats.

4.4.1.2 Marine Dataset

While most of the environmental terms observed for the two different marine based sub-habitats were similar, the ranking of the terms themselves were different and some environmental terms were uniquely observed such as "coral reef" environmental term for coral atoll sample, due to differences in the environment between these two sub-habitats and the variation in taxa abundance that comes with it (Jeffries et al., 2015). Some microbial communities in coral reef systems exist in a symbiotic relationship with coral polyps, playing a role in nutrient cycling as well as assisting in disease resistance for these organisms (Garren & Azam, 2012). Therefore, taxa belonging to this environment are more likely to be observed for coral atoll samples. "Marine bulk water" was uniquely observed for southern ocean samples while being absent for coral atoll samples, due to the environmental characteristic of the ocean waters and the taxa that are prevalent in it.

For all the datasets used for analysis, it was apparent that SEQenv was able to determine intra-habitat differences and patterns even at environmental term level of information.

4.4.2 Per Environment Term Taxa Abundance

As seen in the word clouds for the habitats, certain environmental terms were more strongly observed compared to others. Underpinning this pattern is the taxa abundance, which contributed to their ranking. The Per Environmental Term Taxa Abundance approach was able to provide a more taxa centric explanation of these patterns, which could not be explained solely by SEQenv (Sinclair L, 2016).

Per Environmental Term Taxa Abundance showed distinct patterns of taxa abundances across various environmental terms. Taxa more prevalent in one term were less abundant in another. Certain taxa had low abundances, however depending on the environmental factors these taxa can become more abundant if the conditions are beneficial towards their growth.

4.4.2.1 Sugarcane Dataset

The difference in the abundance of taxa between “rhizosphere” and “soil” terms illustrate that while some taxa were common across different environment, the abundances observed were different.

Acidothermus, which was strongly observed in the "soil" environment term, is a thermophilic, acidophilic, cellulolytic bacterium, prevalent in acidic environments (Mohagheghi, Grohmann, Himmel, Leighton, & Updegraff, 1986), while *Acidobacteriaceae* as observed more in the "forest soil" environment term, is a family of *Acidobacteria* which are ubiquitous in soil environment (Quaiser et al., 2003).

The "garden" environment term was significantly different from other terms in the case of the sequences that contributed to it where "*Pantoea*" was the most abundant taxa observed. It is well known that *Pantoea Spp.* lives in many plant tissues both as commensal and in some cases as pathogens (Pataky, Michener, Freeman, Weinzierl, & Teyker, 2000).

Members of the *Sphingomonas* genus were observed for the environmental term "contaminated soil" and bacteria belonging to this genus is well known to have the ability to degrade chemicals in contaminated soil as it is one of the best known genera for biodegradation of chemical contaminants (Alvarez et al., 2012; S. Schmidt et al., 1992; Ye, Siddiqi, Maccubbin, Kumar, & Sikka, 1995). The most prevalent species of *Sphingomonas* was observed to be an "uncultured marine bacterium". The presence of this bacterium here may be due to this taxon being prevalent in both contaminated soil and marine habitats.

Lastly, while taxa that contributed to the terms such as "contaminated soil" and "waste" were not as abundant as the aforementioned terms like "soil", "forest soil" or "garden", they were nonetheless very important as they provided taxa

abundances under a specific environmental context. Therefore, for studies that may aim towards a specific goal in mind, such as bioremediation, this may help in targeting sequences that come from relevant environments

4.4.2.2 Marine Dataset

Overall, Marine habitats showed an interesting collection of taxa that come from a variety of marine environments. Similar to the sugarcane dataset, while the list of sequences contributing to each environment may seem similar at first, there were exceptions where unique sequences were observed to be more abundant in specific environments. Furthermore, the ranking itself varied across every environmental term. Additionally, similar to the differences observed for "soil" and "forest soil" environmental terms in the sugarcane dataset, "sea" and "seawater" exhibited the same pattern with respect to the taxa observed.

Prochlorococcus, which was observed in multiple environmental terms such as "sea water", "sea" and "ocean" in different abundances, is a very small marine cyanobacteria, which is one of the most abundant photosynthetic organism on the planet (Partensky, Hess, & Vaulot, 1999), while bacteria belonging to *SAR11 clade* are accountable for methane dissolved in the oceans (Carini, White, Campbell, & Giovannoni, 2014). They are cosmopolitan and abundant across marine habitats, particularly *SAR11*, which is the main marine bacterium and was present for most environment terms at different abundances.

Synechococcus is a unicellular cyanobacteria that is prevalent in the marine environment and has been shown to dominate in this system (Jeffries et al., 2015). It was present for the environmental terms "sea" and "sea water", while being absent in top 10 ranked list of taxa for the "ocean" term. *SAR86 Clade*, members of which are aerobic chemoheterotroph (Dupont et al., 2012), and the aforementioned *Prochlorococcus* jointly dominated the "Brine Pool" environment term.

The per environment term taxa abundance provided a more concise and relevant view of the environmental annotations. Linking sequences to environmental terms in such a manner would be more suitable than a list of environmental terms that SEQenv provides. This enhancement significantly improved the analysis capability of SEQenv system and provided a novel approach to contextual, taxa based environmental annotation, which was originally not present in the SEQenv pipeline. Furthermore, the integration developed here enabled a more thorough approach towards 16S rRNA sequence analysis and offers a single pipeline for both taxonomic and environmental annotation of sequences.

4.4.3 Per Taxa Environmental Term Abundance

Following up on per term taxa abundance, similar patterns were observed for per taxa term abundance where certain environmental terms were dominant for specific genus. The per taxa environmental term abundance provided a taxa

centric approach toward environmental annotations and listed the many habitats under which a taxon may be found.

Terms such as "sea" and "soil" are more prevalent due to the limitations associated with the SEQenv pipeline or the meta data for these sequences were not specific enough with respect to the environments they were isolated from.

4.4.3.1 Sugarcane Dataset

In accordance with per term taxa abundance result for "soil" environment term, "soil" dominated the list of terms for *Acidothermus*, which is a thermophilic and acidophilic microbe that is found in acidic environment (Mohagheghi et al., 1986). Other terms such as "forest soil" or "woodland" point towards these environments being favourable to its growth, as it has been observed in samples collected from forest environment (J.-S. Kim et al., 2015; Meng et al., 2013).

Burkholderia occupies a variety of environmental niches (Compant, Nowak, Coenye, Clément, & Ait Barka, 2008) including soil (Janssen, 2006) and some strains of this genus can cause diseases for humans and animals (Coenye & Vandamme, 2003). Furthermore, the bacterium is observed to be prevalent in rhizosphere environment for plants (Caballero-Mellado, Onofre-Lemus, Estrada-de Los Santos, & Martinez-Aguilar, 2007), which may be the reason why the environmental term "rhizosphere" was strongly observed for it as compared to *Acidothermus*. Finally, the presence of the term "sludge" maybe be due to its potential and application for biodegradation (L. Zhang et al., 2013). Overall, this

data supports the widespread distribution in plant rhizosphere of these taxa in multiple niches.

4.4.3.2 Marine Dataset

Prochlorococcus, one of the most abundant organism on the planet (Partensky et al., 1999), is typically observed in oligotrophic oceans where nutrients availability is poor, in contrast to *Synechococcus* that favours nutrient rich environment (Whitton, 2012). Hence terms such as "ocean" and "brine pool" points towards prevalence of *Prochlorococcus* in these environments.

The list of terms for *Synechococcus* includes "reef" and "coast" which are nutrient rich environments compared to oceans. In fact, the bacterium has been observed to be present in high abundance at coral reefs especially during summer time (Moriarty, Pollard, & Hunt, 1985) as well as coastal regions such as the Portuguese coast (Martins, Pereira, Welker, Fastner, & Vasconcelos, 2005).

Overall, the enhancement provided robust data on taxa-specific distribution in different habitats and highlights the usefulness of this approach for delineating the niches potentially occupied by specific taxa, in this case supporting the known distribution of these abundant marine autotrophs, which drive primary production (Christaki, Jacquet, Dolan, Vaulot, & Rassoulzadegan, 1999).

4.5 Conclusion

SEQenv represents a novel method of augmenting sequence analysis with environmental metadata. Given the need for improved analysis, pipelines that can integrate taxonomic and environmental data are becoming increasingly important.

By integrating SEQenv with TaxaSE and extending the functionality through generation of per environment taxa abundance as well as per taxa term abundance data, the improved SEQenv offers unique insights and contributes to the expanding repertoire of next-gen sequence analysis pipelines. This enables the extended pipeline to provide environmental annotations in a variety of contexts.

Furthermore, by directly producing environmental source information for sequences in the dataset, it can greatly help biologists aiming to understand the biogeography of microbes. Given that more and more sequences and genomes are being submitted to the NCBI database, along with associated metadata such as isolation sources, the capabilities of the pipeline would improve in the future.

The combination of enhanced taxonomic annotation, coupled with environmental annotation presents a unique approach to microbial 16S rRNA analysis. The system is capable of accurately annotating environmental information to query sequences and enhancement done to SEQenv, which links taxa to environmental keywords, enhances the applicability of this pipeline. This enhancement would play a greater role in helping ecologists understand the diversity patterns present across diverse habitats and will lead to a holistic approach towards ecological projects.

Chapter 5: Final Conclusion and Future Work

5.1 Conclusion

Given the need for enhanced analysis tools for microbial projects, the overall aim of this thesis was to develop a new system that would fulfil the need of a researcher aiming to investigate amplicon datasets in a more thorough fashion. Annotation of the 16S rRNA gene is the standard approach for taxonomic annotation of bacterial sequences, where sequence similarity determines assignment of taxonomy. However, given that the 16S rRNA gene contains nine regions of variability that can serve to annotate a given sequence better, there is a need to exploit this hypervariable information, as the current methods so far do not account for it.

Additionally, environmental annotation can dramatically enhance our knowledge of microbial world, their niche and distribution. Contextual information, especially environmental data can provide a lot more detail about sequences that may otherwise be left out of the analysis. Given the importance of the roles microbes play in the environment they reside in as well as the various processes they carry out, extracting environmental information about these bacteria would significantly enhance the analysis capability of any system. The most important findings of the study were:

- Evolutionary conservation within the 16S rRNA gene was exploited via Shannon Entropy for taxonomic annotation, providing a novel method for development of a new pipeline. *In-silico* analysis illustrated that the new

weighted scheme built on Shannon entropy is comparable to percentage identity, while showing better performance at higher taxon levels and given that a vast majority of sequences remain uncultured and unknown, this would improve the ability of the system to annotate an unknown sequence much more effectively. The use of a sequence similarity metric that utilizes evolutionary conservation within 16S rRNA gene sequences and consequently improves on annotation of novel sequences at higher taxa level makes this approach different from standard percentage identity based methods.

- A new bioinformatics pipeline, the TaxaSE system, was developed using the Shannon entropy based weighted approach as its foundation. Both TaxaSE and QIIME were used to analyse a large dataset of samples from sugarcane microhabitats, consisting of samples from various habitats. The results showed that the new pipeline was able to generate similar ecologically relevant results for both alpha diversity and beta diversity based analysis. The system can be used for annotating OTUs or perform single sequence annotation at the cost of computational time. Furthermore, as tools were developed to integrate the pipeline within QIIME, researchers can use the new system readily. The OTU independent approach is an alternative to OTU based methods, which can be used to determine taxonomic annotation on a per sequence basis.

- Environmental annotation of microbial sequences was performed using the extended SEQenv pipeline. The results were generated as word clouds,

which help visualising the results much more effectively as well as pie charts to illustrate environmental data more appropriately for sequences. The extension to SEQenv was able to provide a concise, relevant view into the distribution of taxa across different environment terms. The discriminating taxa were ecologically relevant to the specific habitats the new system is able to generate valid environmental annotations. Furthermore, providing environmental terms for sequences in a direct fashion would enable a more thorough and comprehensive approach towards microbial analysis, which integrates niche terms and taxonomic annotations in the same pipelines. Considering that at the moment, taxonomic annotation pipelines do not produce environmental annotations, this approach provides a better picture of bacterial communities in the eco-system, both taxonomically and environmentally.

The combined pipeline, consisting of TaxaSE and expanded SEQenv, provided a single approach for taxonomic annotation of bacterial sequences, enhanced with environmental information. By way of a novel approach of exploiting evolutionary conservation within 16S rRNA gene for taxonomic annotation as well as generating environmental data for these sequences through an extended SEQenv pipeline, a more thorough analysis of ecological projects can be conducted. Thus, the new the pipeline is a novel and sophisticated tool and can greatly augment research seeking to enhance our understanding of microbes and the important roles they play in the environment. Given the need for more thorough analysis, pipelines such as the one described here would be an important addition in a biologist's arsenal of bioinformatics tools.

5.2 Future work

In addition to bacterial 16S rRNA gene sequences, the Shannon entropy approach to quantitatively accessing evolutionary conservation can be applied on other types of sequences where fully aligned databases are available, including archaea 16S rRNA gene sequences, which would require a rebuilding of entropy vectors and few changes in other tools in the pipeline, as well as proteins, as they also have variable and conserved regions. However, this would necessitate larger changes to be done across the pipeline. Overall, this would enable the characterization of sequence divergence in a more effective manner and can perhaps be used to improve taxonomic placement of new sequences as well as develop similar systems for protein annotation. Furthermore, while the TaxaSE pipeline was developed for analyzing amplicon datasets, in future it may be extended to work on whole genome shotgun datasets by extracting short reads which belong to 16S rRNA gene sequences. Various tools are already available that can generate a list of sequences that may be 16S rRNA gene sequences.

Phylogenetic placement algorithms such as pplacer or EPA can augment diversity analysis of microbial community. A combination of both taxonomic and phylogenetic analysis would provide a more comprehensive understanding of the origins of unknown sequences. This would enable better characterization of microbial community composition and may in fact help in determining the content of "microbial dark matter".

Future work on environmental annotations could focus on developing a comprehensive database of sequences and the environments they belong to, which covers all OTUs present in public databases. Chapter 4 highlights the usefulness of this approach within a specific database, which can then be extended to cover a whole range of reference sequences. This can then act as a repository and would be useful in many ecological projects, enabling characterization of microbes on a global level under the context of environments they reside in. While current work was on 16S rRNA gene sequences, SEQenv can run on other sequences as well, such as nucleotide or protein. The extension developed here can be used in the formation of the aforementioned database. Additionally, a multitude of datasets could be analyzed to see how taxonomy is globally partitioned by habitats. Furthermore, if sampling was done over a large time span, environmental annotations at each time step can be generated, which can then elucidate how environmental conditions affected the eco system up to the present day and the resultant microbial community present in the environment, as environmental annotation of a current dataset can only provide a single snapshot. Lastly, the pipeline could be enhanced even more by incorporating extraction of numeric information such as pH, by way of performing text-mining onto research articles as well as dataset metadata and can augment environment annotation by providing environmental variables that can influence the abundance of taxon.

Overall, by enhancing the resolution of annotations and understanding the distribution of taxa across niches, next generation sequencing can realize its potential to understand biodiversity and the underlying mechanisms that generate and sustain it.

Appendix A.

Chapter 4 Table A-1: Top 20 ranked environment terms and associated aggregated values generated for rhizosphere samples from sugarcane dataset

Sugarcane - Rhizosphere		
Rank	Environment Term	Total
1	soil	1144.7871
2	forest soil	337.72467
3	rhizosphere	106.670784
4	forest	94.84745
5	paddy field soil	52.98755
6	pasture	47.228493
7	volcanic field	37.7277
8	peat soil	34.457375
9	rice field	31.99949
10	sediment	31.244991
11	subtropical	30.482468
12	cultivated habitat	27.738499
13	waste	26.490353
14	woodland	21.617014
15	grassland soil	20.789787
16	prairie	19.62367

17	lake	19.454222
18	biofilm	18.735514
19	peat swamp	18.695932
20	field soil	18.084253

Chapter 4 Table A-2: Top 20 ranked environment terms and associated aggregated values generated for soil samples from sugarcane dataset

Sugarcane - Soil		
Rank	Environment Term	Total
1	soil	1124.0039
2	forest soil	260.48883
3	rhizosphere	109.118256
4	paddy field soil	77.9765
5	forest	71.975075
6	pasture	58.9166
7	rice field	55.88218
8	cultivated habitat	49.58408
9	sediment	36.56245
10	subtropical	34.88536
11	red soil	31.142296
12	prairie	29.37707
13	agricultural soil	28.012053
14	field soil	26.450457
15	grassland soil	26.256002
16	biofilm	25.316757

17	peat soil	22.257133
18	volcanic field	20.662825
19	lake	20.261036
20	waste	19.986427

Chapter 4 Table A-3: Top 20 ranked environment terms and associated aggregated values generated for stem samples from sugarcane dataset

Sugarcane - Stem		
Rank	Environment Term	Total
1	soil	617.45306
2	garden	269.33344
3	glacier	59.549744
4	forest soil	52.138023
5	rhizosphere	47.685837
6	biofilm	47.152836
7	ground water	43.154423
8	pasture	28.458822
9	agricultural soil	25.406353
10	mud	22.94665
11	1000047	21.116476
12	sediment	16.962046
13	forest	15.388543
14	waste water	14.882552
15	sea water	14.309931
16	rice field	14.2581625

17	activated sludge	13.9015465
18	lake	13.097423
19	paddy field soil	12.0954685
20	sea	9.209969

Chapter 4 Table A-4: Top 20 ranked environment terms and associated aggregated values generated for root samples from sugarcane dataset

Sugarcane - Root		
Rank	Environment Term	Total
1	soil	1200.9491
2	forest soil	205.77882
3	rhizosphere	158.56563
4	forest	62.163578
5	paddy field soil	39.534786
6	sediment	36.42738
7	pasture	34.83387
8	biofilm	32.352753
9	rice field	31.301132
10	lake	28.31774
11	waste	27.948946
12	volcanic field	27.603746
13	subtropical	25.325792
14	cultivated habitat	24.647427
15	wetland	22.573952
16	field soil	22.43901

17	peat soil	21.702671
18	agricultural soil	19.782148
19	prairie	19.146938
20	grassland soil	17.527967

Chapter 4 Table A-5: Top 20 ranked environment terms and associated aggregated values generated for coral atoll samples from marine dataset

Marine – Coral Atoll		
Rank	Environment Term	Total
1	sea water	697.5574
2	sea	289.85495
3	reef	117.38839
4	ocean	53.56404
5	surface water	40.464886
6	bay	35.168877
7	coast	34.080147
8	brine pool	22.126064
9	tropical	22.05261
10	coral reef	18.642195
11	ocean water	12.49691
12	continental slope	9.513588
13	coastal water body	9.373184
14	mediterranean	9.247752
15	marine bulk water	9.209189
16	447	5.0332727

17	marine habitat	4.3730197
18	coastal water	4.145986
19	lake	4.030988
20	basalt	3.9079504

Chapter 4 Table A-6: Top 20 ranked environment terms and associated aggregated values generated for southern ocean samples from marine dataset

Marine - Southern Ocean		
Rank	Environment Term	Total
1	sea water	425.54196
2	sea	349.8343
3	ocean	152.75682
4	brine pool	99.63226
5	bay	65.48789
6	surface water	33.011257
7	mediterranean	24.882048
8	reef	22.065239
9	marine bulk water	17.266914
10	coast	13.879977
11	447	10.897329
12	lake	10.758537
13	marine habitat	10.70642
14	ocean water	9.395881
15	tropical	9.020766

16	basalt	7.803199
17	continental slope	7.773911
18	coral reef	5.1999903
19	hydrothermal vent	4.946499
20	upwelling	4.451309

Chapter 4 Table A-7: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "soil" in the sugarcane dataset

Sugarcane - Soil Term		
Rank	Taxa	Total
1	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidothermaceae;Acidothermus;	165.0306
2	Bacteria;Cyanobacteria;Chloroplast;uncultured eukaryote	127
3	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;uncultured bacterium	96.839264
4	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;	89.18057
5	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;	79.01076
6	Bacteria;Cyanobacteria;Chloroplast;Lolium perenne	70
7	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;	57.734806
8	Bacteria;Actinobacteria;Thermoleophilia;Gaiellales;uncultured;	56.530354
9	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiales Incertae Sedis;Rhizomicrobium;	54.74491

10	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Bradyrhizobiaceae;Bradyrhizobium;uncultured bacterium	53.16674
11	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured Acidobacteria bacterium	50.429787
12	Bacteria;Cyanobacteria;Chloroplast;	44.5
13	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;uncultured;	43.65489
14	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured bacterium	43.36269
15	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;DA111;	42.765457
16	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;uncultured;	38.10091
17	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Massilia;uncultured bacterium	36.15
18	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;	35.50737
19	Bacteria;Cyanobacteria;Chloroplast;Hordeum vulgare subsp. vulgare (domesticated barley)	35
20	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas;uncultured bacterium	34.26905

Chapter 4 Table A-8: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "forest soil" in the sugarcane dataset

Sugarcane - Forest Soil Term

Rank	Taxa	Total
1	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;	56.819992
2	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Variibacter;uncultured bacterium	38.81915
3	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured bacterium	33.71271
4	Bacteria;Acidobacteria;Acidobacteria;Subgroup 2;	33.70088
5	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidothermaceae;Acidothermus;	31.116476
6	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;uncultured bacterium	22.848839
7	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiales Incertae Sedis;Rhizomicrobium;	22.621468
8	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;	21.82482
9	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;uncultured bacterium	21.39969
10	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured Acidobacteria bacterium	18.592148
11	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;	17.27621
12	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadales Incertae Sedis;Acidibacter;	14.721128

13	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;Chitinophaga;	13.992879
14	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Variibacter;	13.323421
15	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;uncultured;	12.74283
16	Bacteria;Acidobacteria;Acidobacteria;Subgroup 2;uncultured bacterium	12.4988
17	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;uncultured;	12.386102
18	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Telmatobacter;uncultured bacterium	11.883281
19	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;alphaI cluster;	11.243621
20	Bacteria;Actinobacteria;Thermoleophilia;Gaiellales;uncultured;	10.9536

Chapter 4 Table A-9: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "rhizosphere" in the sugarcane dataset

Sugarcane - Rhizosphere Term		
Rank	Taxa	Total
1	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;	34.710728
2	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;uncultured bacterium	11.230921

3	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Sphingobacteriaceae;Mucilaginibacter;	11.0611
4	Bacteria;Actinobacteria;Actinobacteria;Catenulisporales;Catenulisporaceae;Catenulispora;Catenulispora sp. Neo1	10.61667
5	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;	8.643299
6	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Dyella;	8.116631
7	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;Burkholderia kururiensis subsp. kururiensis	8
8	Bacteria;Actinobacteria;Actinobacteria;Catenulisporales;Catenulisporaceae;Catenulispora;	6.9712152
9	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Burkholderia;Burkholderia sp. USM	6.5
10	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Acidicapsa;	6.49994
11	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Rhodanobacter;	6.1726
12	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Mesorhizobium;Mesorhizobium plurifarium	6
13	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Acidicapsa;Acidicapsa sp. CE1	5.91661
14	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured Acidobacteria bacterium	5.7277694

15	Bacteria;Planctomycetes;Phycisphaerae;WD2101 soil group;uncultured bacterium	5.365153
16	Bacteria;Actinobacteria;Thermoleophilia;Gaiellales;uncultured;	5.256879
17	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;DA111;	5.1761904
18	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadales Incertae Sedis;Acidibacter;	5.028715
19	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Bradyrhizobiaceae;Bradyrhizobium;uncultured proteobacterium	5
20	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;	4.6012993

Chapter 4 Table A-10: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "garden" in the sugarcane dataset

Sugarcane - Garden Term		
Rank	Taxa	Total
1	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea;uncultured bacterium	160.33334
2	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea;	58.666687
3	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea;Pantoea stewartii	33
4	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea;Pantoea sp. NG8	14.666738

5	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Stenotrophomonas;uncultured bacterium	4
6	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Acinetobacter;Acinetobacter sp. C008	3
7	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Tatumella;uncultured bacterium	2
8	Bacteria;Actinobacteria;Actinobacteria;Streptomycetales;Streptomycetaceae;Streptacidiphilus;Streptacidiphilus sp. 5-20	1.91665
9	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea;gamma proteobacterium symbiont of Plautia stali	1
10	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Enterobacter;	0.66666
11	Bacteria;Actinobacteria;Actinobacteria;Streptomycetales;Streptomycetaceae;Streptacidiphilus;	0.58333004
12	Bacteria;Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;uncultured;	0.555989
13	Bacteria;Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;uncultured;uncultured Gemmatimonas sp.	0.54545397
14	Bacteria;Proteobacteria;Deltaproteobacteria;GR-WP33-30;	0.5
15	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;uncultured bacterium	0.4
16	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Mitochondria;	0.33333

17	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;	0.29167
18	Bacteria;Actinobacteria;Actinobacteria;Pseudonocardiales;Pseudonocardiaceae;Pseudonocardia;	0.25
19	Bacteria;Acidobacteria;Acidobacteria;Subgroup 2;	0.181818
20	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Acidobacterium;uncultured bacterium	0.16667

Chapter 4 Table A-11: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "forest" in the sugarcane dataset

Sugarcane - Forest Term		
Rank	Taxa	Total
1	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidothermaceae;Acidothermus;	21.23047
2	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;uncultured;	10.50042
3	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Acidicapsa;	7.8333206
4	Bacteria;Proteobacteria;Betaproteobacteria;SC-I-84;	5.7622795
5	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;uncultured;	5.54168
6	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;uncultured;uncultured bacterium	4.95

7	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;	4.6916895
8	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;uncultured;	4.266098
9	Bacteria;Actinobacteria;Actinobacteria;Micromonosporales;Micromonosporaceae;Actinocatenispora;	4.1666603
10	Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Polyangiaceae;Sorangium;	4.0830398
11	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;	3.991361
12	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured Acidobacteria bacterium	3.694446
13	Bacteria;Acidobacteria;Holophagae;Subgroup 7;	3.67977
14	Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Haliangiaceae;Haliangium;	3.65834
15	Bacteria;Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;uncultured;	3.4706302
16	Bacteria;Acidobacteria;Acidobacteria;Subgroup 2;	3.4128518
17	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;uncultured;	3.31984
18	Bacteria;Planctomycetes;Phycisphaerae;WD2101 soil group;uncultured bacterium	3.2576318
19	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillales Incertae Sedis;Reyranella;	3.19286

20	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadales Incertae Sedis;Acidibacter;	3.1626148

Chapter 4 Table A-12: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "paddy field soil" in the sugarcane dataset

Sugarcane - Paddy Field Soil Term		
Rank	Taxa	Total
1	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;	11.78251
2	Bacteria;Actinobacteria;Thermoleophilia;Gaiellales;unculture d;	7.499279
3	Bacteria;Actinobacteria;Thermoleophilia;Gaiellales;unculture d;uncultured bacterium	6.7321906
4	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidothermaceae;Acidothermus;	6.449308
5	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadales Incertae Sedis;Acidibacter;	5.440577
6	Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;uncultured;	4.68334
7	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiales Incertae Sedis;Rhizomicrobium;	4.5500298
8	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;uncultured;	4.46904
9	Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacteraceae;Anaeromyxobacter;	4.30953

10	Bacteria;Proteobacteria;Betaproteobacteria;SC-I-84;	4.273829
11	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;Chitinophaga;	3.5833201
12	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured Acidobacteria bacterium	3.5
13	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured bacterium	3.26786
14	Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Haliangiaceae;Haliangium;	3.26469
15	Bacteria;Proteobacteria;Betaproteobacteria;Nitrosomonadales;Nitrosomonadaceae;uncultured;	3.12738
16	Bacteria;Acidobacteria;Acidobacteria;Subgroup 3;Unknown Family;Candidatus Solibacter;	3.058124
17	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Planctomyces;	3.0485947
18	Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;Mycobacteriaceae;Mycobacterium;Mycobacterium sp. QIA-36	3
19	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfuromonadales;Geobacteraceae;Geobacter;	2.8666701
20	Bacteria;Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;uncultured;	2.8476589

Chapter 4 Table A-13: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "contaminated soil" in the sugarcane dataset

Sugarcane - Contaminated Soil Term		
Rank	Taxa	Total
1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas;uncultured marine bacterium	2.66666
2	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas;uncultured bacterium	2
3	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Undibacterium;bacterium PH2(2012)	1.9999801
4	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas;uncultured Kaistobacter sp.	1.55
5	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas;	1.45237
6	Bacteria;Actinobacteria;Actinobacteria;Kineosporiales;Kineosporiaceae;Quadrisphaera;	1.33333
7	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidothermaceae;Acidothermus;uncultured bacterium	1.2
8	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas;	1
9	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;uncultured;uncultured Acetobacteraceae bacterium	0.95

10	Bacteria;Verrucomicrobia;OPB35 soil group;uncultured Verrucomicrobia subdivision 3 bacterium	0.85716
11	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;uncultured;	0.8373
12	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Hyphomicrobium;uncultured bacterium	0.81667
13	Bacteria;Chloroflexi;KD4-96;	0.79222
14	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;DA111;	0.750003
15	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Dyella;	0.7143
16	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Rhizobium;	0.66666996
17	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Rhodanobacter;	0.64286
18	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas;unidentified marine bacterioplankton	0.6
19	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Beijerinckia;Beijerinckia doebereinerae	0.58334
20	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Variibacter;	0.58333004

Chapter 4 Table A-14: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "waste" in the sugarcane dataset

Sugarcane - Waste Term		
Rank	Taxa	Total
1	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidothermaceae;Acidothermus;	15.420239
2	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;uncultured Acidobacteria bacterium	7.2555456
3	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;uncultured;	6.1944404
4	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);Granulicella;	3.7135706
5	Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae (Subgroup 1);uncultured;	3.400835
6	Bacteria;Bacteroidetes;Sphingobacteriia;Sphingobacteriales;Chitinophagaceae;Chitinophaga;	3.2761998
7	Bacteria;Verrucomicrobia;Spartobacteria;Chthoniobacterales;Chthoniobacteraceae;Chthoniobacter;	3.0954542
8	Bacteria;Actinobacteria;Actinobacteria;Glycomycetales;Glycomycetaceae;Glycomyces;Glycomyces algeriensis	1.9166502
9	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;uncultured;	1.85
10	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadales Incertae Sedis;Acidibacter;	1.652859

11	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales; Planctomycetaceae;Planctomyces;	1.43571
12	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales ;Rhodospirillaceae;uncultured;	1.34286
13	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonad ales;Pseudomonadaceae;Pseudomonas;Pseudomonas oryzihabitans	1.33333
14	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales ;Acetobacteraceae;uncultured;uncultured bacterium	1.33332
15	Bacteria;Actinobacteria;Actinobacteria;Frankiales;Acidotherm aceae;Acidothermus;uncultured bacterium	1.2833301
16	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonad ales;Pseudomonadaceae;Pseudomonas;Pseudomonas sp. PPF-2	1
17	Bacteria;Actinobacteria;Actinobacteria;Glycomycetales;Glyco mycetaceae;Glycomyces;	0.99999
18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhiz obiales Incertae Sedis;Rhizomicrobium;	0.97499996
19	Bacteria;Actinobacteria;Thermoleophilia;Solirubrobacterales; 480-2;	0.78572
20	Bacteria;WD272;uncultured bacterium	0.76668

Chapter 4 Table A-15: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "seawater" in the marine dataset

Marine - Seawater Term	

Rank	Taxa	Total
1	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus;	466.62323
2	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1;	122.371796
3	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;	120.725784
4	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus;	101.50537
5	Bacteria;Cyanobacteria;Chloroplast;	67.342094
6	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured;	33.00002
7	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;	31.138363
8	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2;	18.230541
9	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SAR116 clade;	16.58454
10	Bacteria;Marinimicrobia (SAR406 clade);	14.2154875
11	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;	11.03702
12	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Oceanospirillaceae;Pseudospirillum;	10
13	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OM1 clade;Candidatus Actinomarina;	8.649969
14	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;NS9 marine group;	8.299818

15	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Cryomorphaceae;Owenweeksia;	8.2142
16	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);	7.9523597
17	Bacteria;Chloroflexi;SAR202 clade;	6.2666206
18	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;S25-593;	5.93333
19	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS5 marine group;	5.57857
20	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 4;	5.39285

Chapter 4 Table A-16: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "sea" in the marine dataset

Marine - Sea Term		
Rank	Taxa	Total
1	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1;	157.70627
2	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus;	126.87572
3	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;	104.731735
4	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus;	79.78276
5	Bacteria;Cyanobacteria;Chloroplast;	50.740696
6	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;	18.02246

7	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;	13.54829
8	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2;	12.840851
9	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured;	11.341699
10	Bacteria;Marinimicrobia (SAR406 clade);	10.73333
11	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);	7.250019
12	Bacteria;Chloroflexi;SAR202 clade;	6.51108
13	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;NS9 marine group;	5.1142898
14	Bacteria;Proteobacteria;Alphaproteobacteria;OCS116 clade;	3.3250003
15	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OM1 clade;Candidatus Actinomarina;	2.9166698
16	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SAR116 clade;	2.90237
17	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS4 marine group;	2.25077
18	Bacteria;Proteobacteria;Gammaproteobacteria;Cellvibrionales;Porticoccaceae;SAR92 clade;	2
19	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS5 marine group;	1.73571
20	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;Defluviicoccus;	1.6363701

Chapter 4 Table A-17: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "ocean" in the marine dataset

Marine - Ocean Term		
Rank	Taxa	Total
1	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1;	36.824417
2	Bacteria;Cyanobacteria;Chloroplast;	33.787674
3	Bacteria;Marinimicrobia (SAR406 clade);	21.18491
4	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;	17.01603
5	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;	12.90271
6	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus;	12.150019
7	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;NS9 marine group;	10.420288
8	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfuromonadales;GR-WP33-58;	5.80952
9	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured;	5.5833693
10	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2;	5.3571496
11	Bacteria;Proteobacteria;Alphaproteobacteria;OCS116 clade;	5.2666693
12	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 4;	3.9643
13	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SAR116 clade;	3.36666

14	Bacteria;Chloroflexi;SAR202 clade;	3.1110795
15	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 3;	2.91667
16	Bacteria;Proteobacteria;Gammaproteobacteria;E01-9C-26 marine group;	2
17	Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Marinoscillum;	1.861093
18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;uncultured;	1.5
19	Bacteria;Verrucomicrobia;Opitutae;MB11C04 marine group;	1.4000001
20	Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;OM27 clade;	1.25

Chapter 4 Table A-18: Top 15 ranked list of taxa and associated aggregate values observed for the environmental term "reef" in the marine dataset

Marine - Reef Term		
Rank	Taxa	Total
1	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus;	103.707756
2	Bacteria;Cyanobacteria;Chloroplast;	16.772291
3	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;	6.4448986
4	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Cryomorphaceae;Owenweeksia;	2.8571992
5	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1;	2.2666702

6	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;	1.626193
7	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2;	1.5444499
8	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;NS9 marine group;	1.4166899
9	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS5 marine group;	1
10	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured;	0.54166996
11	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;	0.36111
12	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;Sva0996 marine group;	0.28572
13	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Cryomorphaceae;	0.26786
14	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;	0.25
15	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS4 marine group;	0.11111

Chapter 4 Table A-19: Top 8 ranked list of taxa and associated aggregate values observed for the environmental term "brine pool" in the marine dataset

Marine - Brine Pool Term		
Rank	Taxa	Total

1	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;	59.449463
2	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus;	57.52571
3	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;	2.16666
4	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);	1.49999
5	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1;	0.516663
6	Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Alteromonas;	0.33333
7	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured;	0.16667
8	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfuromonadales;GR-WP33-58;	0.1

Chapter 4 Table A-20: Top 20 ranked list of taxa and associated aggregate values observed for the environmental term "bay" in the marine dataset

Marine - Bay		
Rank	Taxa	Total
1	Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade;	26.377739
2	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1;	16.227812
3	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OM1 clade;Candidatus Actinomarina;	7.4999695

4	Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured;	5.5583706
5	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;NS9 marine group;	4.2942786
6	Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus;	4.16665
7	Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SAR116 clade;	3.5012002
8	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfuromonadales;GR-WP33-58;	2.95237
9	Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B);	2.8631098
10	Bacteria;Cyanobacteria;Chloroplast;	2.6048813
11	Bacteria;Proteobacteria;Gammaproteobacteria;Cellvibrionales;Halieaceae;OM60(NOR5) clade;	2.5138798
12	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 3;	2.375
13	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;NS2b marine group;	2
14	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2;	1.70398
15	Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;OM27 clade;	1.58333
16	Bacteria;Marinimicrobia (SAR406 clade);	1.5000001
17	Bacteria;Proteobacteria;Alphaproteobacteria;OCS116 clade;	1.10834

18	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Cryomorphaceae;Owenweeksia;	1
19	Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;	0.974678
20	Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;Sva0996 marine group;	0.93506

References

Alimetrics. DNA SEQUENCE ANALYSIS. Retrieved from http://www.alimetrics.net/en/index.php/dna-sequence-analysis

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Alvarez, A., Benimeli, C. S., Saez, J. M., Fuentes, M. S., Cuozzo, S. A., Polti, M. A., & Amoroso, M. J. (2012). Bacterial Bio-Resources for Remediation of Hexachlorocyclohexane. *International Journal of Molecular Sciences, 13*(11), 15086-15106. doi:10.3390/ijms131115086

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews, 59*(1), 143-169.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32-46. doi:DOI 10.1111/j.1442-9993.2001.01070.pp.x

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PLoS One, 7*(10), e46679.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., . . . Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics, 9*, 75. doi:10.1186/1471-2164-9-75

Baas-Becking, L. G. M. (1934). *Geobiologie; of inleiding tot de milieukunde*: WP Van Stockum & Zoon NV.

Baker, G. C., Gaffar, S., Cowan, D. A., & Suharto, A. R. (2001). Bacterial community analysis of Indonesian hot springs. *FEMS Microbiology Letters, 200*(1), 103-109. doi:10.1111/j.1574-6968.2001.tb10700.x

Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods, 55*(3), 541-555.

Balvočiūtė, M., & Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genomics, 18*(2), 114.

Barriuso, J., Valverde, J. R., & Mellado, R. P. (2011). Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics, 12*. doi:10.1186/1471-2105-12-473

Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research, 40*(Database issue), D48-53. doi:10.1093/nar/gkr1202

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology, 60*(3), 291-302.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., & Galaxy, T. (2010). Manipulation of FASTQ data with Galaxy. *Bioinformatics, 26*(14), 1783-1785. doi:10.1093/bioinformatics/btq281

Broughton, L., & Gross, K. (2000). Patterns of diversity in plant and soil microbial communities along a productivity gradient in a Michigan old-field. *Oecologia, 125*(3), 420-427.

Brown, M. P. (2000). *Small subunit ribosomal RNA modeling using stochastic context-free grammars.* Paper presented at the ISMB.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*(1), 59-60.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., & Consortium, E. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics, 4*(1), 43. doi:10.1186/2041-1480-4-43

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., & Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics, 7*, 57. doi:10.1186/s13326-016-0097-6

Caballero-Mellado, J., Onofre-Lemus, J., Estrada-de Los Santos, P., & Martinez-Aguilar, L. (2007). The tomato rhizosphere, an environment rich in nitrogen-fixing Burkholderia species with capabilities of interest for agriculture and bioremediation. *Applied and Environmental Microbiology, 73*(16), 5308-5319. doi:10.1128/aem.00324-07

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335-336. doi:10.1038/nmeth.f.303

Carini, P., White, A. E., Campbell, E. O., & Giovannoni, S. J. (2014). Methane production by phosphate-starved SAR11 chemoheterotrophic marine bacteria. *Nature communications, 5*.

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods, 69*(2), 330-339. doi:10.1016/j.mimet.2007.02.005

Chapman, R. W., Robalino, J., & F. Trent III, H. (2006). EcoGenomics: analysis of complex systems via fractal geometry. *Integrative and Comparative Biology, 46*(6), 902-911.

Christaki, U., Jacquet, S., Dolan, J. R., Vaulot, D., & Rassoulzadegan, F. (1999). Growth and grazing on Prochlorococcus and Synechococcus by two marine ciliates. *Limnology and Oceanography, 44*(1), 52-61.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., . . . O'Toole, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature, 488*(7410), 178-184. doi:10.1038/nature11319

CLARKE, K. R. (1993). Non ‐ parametric multivariate analyses of changes in community structure. *Australian journal of ecology, 18*(1), 117-143.

Cloutier, D. D., Alm, E. W., & McLellan, S. L. (2015). Influence of land use, nutrients, and geography on microbial communities and fecal indicator abundance at Lake Michigan beaches. *Applied and environmental microbiology, 81*(15), 4904-4913.

Coenye, T., & Vandamme, P. (2003). Diversity and significance of Burkholderia species occupying diverse ecological niches. *Environmental Microbiology, 5*(9), 719-729.

Coit, P., & Sawalha, A. H. (2016). The human microbiome in rheumatic autoimmune diseases: A comprehensive review. *Clinical Immunology, 170*, 70-79. doi:10.1016/j.clim.2016.07.026

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., . . . Tiedje, J. M. (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research, 33*. doi:10.1093/nar/gki038

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., . . . Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research, 42*(Database issue), D633-642. doi:10.1093/nar/gkt1244

Compant, S., Nowak, J., Coenye, T., Clément, C., & Ait Barka, E. (2008). Diversity and occurrence of Burkholderia spp. in the natural environment. *FEMS Microbiology Reviews, 32*(4), 607-626. doi:10.1111/j.1574-6976.2008.00113.x

Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics, 11*(1), 1-6. doi:10.1186/1471-2105-11-485

Danhorn, T., & Fuqua, C. (2007). Biofilm formation by plant-associated bacteria. *Annual Review of Microbiology, 61*, 401-422. doi:10.1146/annurev.micro.61.080706.093316

De Wit, R., & Bouvier, T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology, 8*(4), 755-758. doi:10.1111/j.1462-2920.2006.01017.x

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., . . . Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology, 72*(7), 5069-5072. doi:10.1128/aem.03006-05

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., . . . Rohwer, F. (2008). Functional metagenomic profiling of nine biomes. *Nature, 452*(7187), 629-632. doi:http://www.nature.com/nature/journal/v452/n7187/suppinfo/nature06810_S1.html

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research, 36*(16), e105-e105.

Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J. P., & Raoult, D. (2000). 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *Journal of Clinical Microbiology, 38*(10), 3623-3630.

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J., Richter, R. A., Valas, R., . . . Haft, D. H. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME journal, 6*(6), 1186-1199.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461. doi:10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods, 10*(10), 996-998. doi:10.1038/nmeth.2604

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics, 27*(16), 2194-2200. doi:10.1093/bioinformatics/btr381

Eren, A. M., Borisy, G. G., Huse, S. M., & Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Science U S A, 111*(28), E2875-2884. doi:10.1073/pnas.1409644111

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution, 4*(12). doi:10.1111/2041-210X.12114

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2015). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME journal, 9*(4), 968-979.

Eren, A. M., Zozaya, M., Taylor, C. M., Dowd, S. E., Martin, D. H., & Ferris, M. J. (2011). Exploring the diversity of Gardnerella vaginalis in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS One, 6*(10), e26732.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research, 8*(3), 175-185.

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science, 320*(5879), 1034-1039. doi:10.1126/science.1153213

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:10.1016/j.patrec.2005.10.010

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., . . . Wooley, J. (2011). The Genomic Standards Consortium. *PLoS Biology, 9*(6), e1001088. doi:10.1371/journal.pbio.1001088

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., . . . Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology, 26*(5), 541-547. doi:http://www.nature.com/nbt/journal/v26/n5/suppinfo/nbt1360_S1.html

Fierer, N., Bradford, M. A., & Jackson, R. B. (2007). Toward an ecological classification of soil bacteria. *Ecology, 88*(6), 1354-1364.

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Science U S A, 103*(3), 626-631. doi:10.1073/pnas.0507535103

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME Journal, 6*(5), 1007-1017. doi:10.1038/ismej.2011.159

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., . . . Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences, 109*(52), 21390-21395. doi:10.1073/pnas.1215210110

Fox, G. E., Wisotzkey, J. D., & Jurtshuk, P., Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic Bacteriology, 42*(1), 166-170. doi:10.1099/00207713-42-1-166

Fredriksson, N. J., Hermansson, M., & Wilen, B. M. (2013). The choice of PCR primers has great impact on assessments of bacterial community diversity and dynamics in a wastewater treatment plant. *PLoS One, 8*(10), e76431. doi:10.1371/journal.pone.0076431

Garbeva, P., Veen, J. A. v., & Elsas, J. D. v. (2004). MICROBIAL DIVERSITY IN SOIL: Selection of Microbial Populations by Plant and Soil Type and Implications for Disease Suppressiveness. *Annual Review of Phytopathology, 42*(1), 243-270. doi:doi:10.1146/annurev.phyto.42.012604.135455

Garren, M., & Azam, F. (2012). New directions in coral reef microbial ecology. *Environmental Microbiology, 14*(4), 833-844. doi:10.1111/j.1462-2920.2011.02597.x

The GenBank Submissions Handbook [Internet]. (2011-). Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK51157/

Ghiglione, J. F., Galand, P. E., Pommier, T., Pedros-Alio, C., Maas, E. W., Bakker, K., . . . Murray, A. E. (2012). Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Science U S A, 109*(43), 17633-17638. doi:10.1073/pnas.1208160109

Gibbons, S. M., & Gilbert, J. A. (2015). Microbial diversity — exploration of natural ecosystems and microbiomes. *Current Opinion in Genetics & Development, 35*, 66-72. doi:http://dx.doi.org/10.1016/j.gde.2015.10.003

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology, 12*, 69. doi:10.1186/s12915-014-0069-1

Gouda, S., Das, G., Sen, S. K., Shin, H. S., & Patra, J. K. (2016). Endophytes: A Treasure House of Bioactive Compounds of Medicinal Importance. *Frontiers in Microbiology, 7*, 1538. doi:10.3389/fmicb.2016.01538

Grosskopf, R., Janssen, P. H., & Liesack, W. (1998). Diversity and structure of the methanogenic community in anoxic rice paddy soil microcosms as examined by cultivation and direct 16S rRNA gene sequence retrieval. *Applied and environmental microbiology, 64*(3), 960-969.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., . . . Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research, 21*(3), 494-504. doi:10.1101/gr.112730.110

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews, 68*(4), 669-685. doi:10.1128/MMBR.68.4.669-685.2004

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., & Martiny, J. B. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology, 10*(7), 497-506. doi:10.1038/nrmicro2795

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., . . . Zhou, H.-W. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome, 3*, 20. doi:10.1186/s40168-015-0081-x

Hooper, L. V., Littman, D. R., & Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science (New York, N.Y.), 336*(6086), 1268-1273. doi:10.1126/science.1223490

Hsu, W., Han, S. X., Arnold, C. W., Bui, A. A., & Enzmann, D. R. (2016). A data-driven approach for quality assessment of radiologic interpretations. *Journal of American Medical Informatics Associations, 23*(e1), e152-156. doi:10.1093/jamia/ocv161

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology, 3*(2), REVIEWS0003.

Hugenholtz, P., Pitulle, C., Hershberger, K. L., & Pace, N. R. (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *Journal of Bacteriology, 180*(2), 366-376.

Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207-214. doi:10.1038/nature11234

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research, 17*(3), 377-386. doi:10.1101/gr.5969107

Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., & Schuster, S. C. (2009). Methods for comparative metagenomics. *BMC Bioinformatics, 10 Suppl 1*, S12. doi:10.1186/1471-2105-10-S1-S12

Ijaz, A. Z. (2017). Collection of datasets containing the TaxaSE bacterial taxonomic annotation pipeline, SILVA insilico datasets and Illumina sequencing data from sugarcane bacterial (16S) including subhabitats from soil,

rhizosphere, stem and root. Retrieved from http://hie-pub.westernsydney.edu.au/6a603c5e-35d6-11e7-b329-525400daae48/

Imam, J., Singh, P. K., & Shukla, P. (2016). Plant Microbe Interactions in Post Genomic Era: Perspectives and Applications. *Frontiers in Microbiology, 7*, 1488. doi:10.3389/fmicb.2016.01488

Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and environmental microbiology, 72*(3), 1719-1728.

Jeffries, T. C., Ostrowski, M., Williams, R. B., Xie, C., Jensen, R. M., Grzymski, J. J., . . . Lauro, F. M. (2015). Spatially extensive microbial biogeography of the Indian Ocean provides insights into the unique community structure of a pristine coral atoll. *Scientific Reports, 5*, 15383. doi:10.1038/srep15383

Jeffries, T. C., Seymour, J. R., Gilbert, J. A., Dinsdale, E. A., Newton, K., Leterme, S. S., . . . Mitchell, J. G. (2011). Substrate type determines metagenomic profiles from diverse chemical habitats. *PLoS One, 6*(9), e25173. doi:10.1371/journal.pone.0025173

Jiang, L., Zhang, J., Xuan, P., & Zou, Q. (2016). BP Neural Network Could Help Improve Pre-miRNA Identification in Various Species. *Biomed Research International, 2016*, 9565689. doi:10.1155/2016/9565689

Ke, P. J., & Miki, T. (2015). Incorporating the soil environment and microbial community into plant competition theory. *Frontiers in Microbiology, 6*, 1066. doi:10.3389/fmicb.2015.01066

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research, 12*(4), 656-664. doi:10.1101/gr.229202. Article published online before March 2002

Kim, J.-S., Lee, K. C., Kim, D.-S., Ko, S.-H., Jung, M.-Y., Rhee, S.-K., & Lee, J.-S. (2015). Pyrosequencing analysis of a bacterial community associated with lava-formed soil from the Gotjawal forest in Jeju, Korea. *MicrobiologyOpen, 4*(2), 301-312. doi:10.1002/mbo3.238

Kim, M., Lee, K. H., Yoon, S. W., Kim, B. S., Chun, J., & Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & Informatics, 11*(3), 102-113. doi:10.5808/GI.2013.11.3.102

Kim, M., Morrison, M., & Yu, Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods, 84*(1), 81-87. doi:http://dx.doi.org/10.1016/j.mimet.2010.10.020

Kirk, J. L., Beaudette, L. A., Hart, M., Moutoglis, P., Klironomos, J. N., Lee, H., & Trevors, J. T. (2004). Methods of studying soil microbial diversity. *Journal of Microbiological Methods, 58*(2), 169-188. doi:http://dx.doi.org/10.1016/j.mimet.2004.04.006

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., . . . Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature Methods, 8*(9), 761-763. doi:10.1038/nmeth.1650

Koch, A. L. (2001). Oligotrophs versus copiotrophs. *BioEssays, 23*(7), 657-661. doi:10.1002/bies.1091

Kostadinov, I. (2011). *Marine Metagenomics: From high-throughput data to ecogenomic interpretation.* Bremen, Jacobs Univ., Diss., 2011.

Kuenen, J. G. (1983). The Role of Specialists and Generalists in Microbial Population Interactions *Foundations of Biochemical Engineering* (Vol. 207, pp. 229-251): AMERICAN CHEMICAL SOCIETY.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Lanzen, A., Jorgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., . . . Urich, T. (2012). CREST--classification resources for environmental sequence tags. *PLoS One, 7*(11), e49334. doi:10.1371/journal.pone.0049334

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution, 21*(6), 1095-1109.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification, 3*, 1-8. doi:10.1016/j.bdq.2015.02.001

Lennon, J. T., & Jones, S. E. (2011). Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology, 9*(2), 119-130. doi:10.1038/nrmicro2504

Li, H. (2017). Toolkit for processing sequences in FASTA/Q formats. Retrieved from https://github.com/lh3/seqtk

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., . . . Acinas, S. G. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and

structure of microbial communities. *Environmental Microbiology, 16*(9), 2659-2671. doi:10.1111/1462-2920.12250

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology, 30*(5), 434-439.

Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C., & Glöckner, F. O. (2006). Megx.net—database resources for marine ecological genomics. *Nucleic acids research, 34*(Database issue), D390-D393. doi:10.1093/nar/gkj070

Lozupone, C. A., & Knight, R. (2007). Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences, 104*(27), 11436-11440. doi:10.1073/pnas.0611525104

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, . . . Schleifer, K. H. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research, 32*(4), 1363-1371. doi:10.1093/nar/gkh293

Mackelprang, R., Saleska, S. R., Jacobsen, C. S., Jansson, J. K., & Taş, N. (2016). Permafrost Meta-Omics and Climate Change. *Annual Review of Earth and Planetary Sciences, 44*(1), 439-462. doi:doi:10.1146/annurev-earth-060614-105126

Magoc, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957-2963. doi:10.1093/bioinformatics/btr507

Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., . . . Relman, D. A. (2007). Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human

mouth. *Proceedings of the National Academy of Sciences, 104*(29), 11889-11894.

Martins, R., Pereira, P., Welker, M., Fastner, J., & Vasconcelos, V. M. (2005). Toxicity of culturable cyanobacteria strains isolated from the Portuguese coast. *Toxicon, 46*(4), 454-464. doi:http://dx.doi.org/10.1016/j.toxicon.2005.06.010

Martiny, J. B., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., . . . Staley, J. T. (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology, 4*(2), 102-112. doi:10.1038/nrmicro1341

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics, 13*, 31. doi:10.1186/1471-2105-13-31

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S., Dubinsky, E. A., Fortney, J. L., . . . Jansson, J. K. (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME Journal, 6*(9), 1715-1727. doi:10.1038/ismej.2012.59

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics, 11*(1), 538.

McCalley, C. K., Woodcroft, B. J., Hodgkins, S. B., Wehr, R. A., Kim, E.-H., Mondav, R., . . . Saleska, S. R. (2014). Methane dynamics regulated by microbial community response to permafrost thaw. *Nature, 514*(7523), 478-481. doi:10.1038/nature13798

http://www.nature.com/nature/journal/v514/n7523/abs/nature13798.html - supplementary-information

McInerney, J. O., Wilkinson, M., Patching, J. W., Embley, T. M., & Powell, R. (1995). Recovery and phylogenetic analysis of novel archaeal rRNA sequences from a deep-sea deposit feeder. *Applied and Environmental Microbiology, 61*(4), 1646-1648.

Meng, H., Li, K., Nie, M., Wan, J.-R., Quan, Z.-X., Fang, C.-M., . . . Li, B. (2013). Responses of bacterial and fungal communities to an elevation gradient in a subtropical montane forest of China. *Applied Microbiology and Biotechnology, 97*(5), 2219-2230. doi:10.1007/s00253-012-4063-7

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., . . . White, O. (2012). A framework for human microbiome research. *Nature, 486*(7402), 215-221. doi:10.1038/nature11209

Mielczarek, M., & Szyda, J. (2016). Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics, 57*(1), 71-79. doi:10.1007/s13353-015-0292-7

Miguel, P. S. B., de Oliveira, M. N. V., Delvaux, J. C., de Jesus, G. L., Borges, A. C., Tótola, M. R., . . . Costa, M. D. (2016). Diversity and distribution of the endophytic bacterial community at different stages of Eucalyptus growth. *Antonie van Leeuwenhoek, 109*(6), 755-771. doi:10.1007/s10482-016-0676-7

Mohagheghi, A., Grohmann, K., Himmel, M., Leighton, L., & Updegraff, D. (1986). Isolation and characterization of Acidothermus cellulolyticus gen. nov., sp. nov., a new genus of thermophilic, acidophilic, cellulolytic bacteria. *International Journal of Systematic and Evolutionary Microbiology, 36*(3), 435-443.

Monard, C., Gantner, S., Bertilsson, S., Hallin, S., & Stenlid, J. (2016). Habitat generalists and specialists in microbial communities across a terrestrial-freshwater gradient. *Scientific Reports, 6*, 37719. doi:10.1038/srep37719

http://www.nature.com/articles/srep37719 - supplementary-information

Morgan, X. C., & Huttenhower, C. (2014). Meta'omic analytic techniques for studying the intestinal microbiome. *Gastroenterology, 146*(6), 1437-1448 e1431. doi:10.1053/j.gastro.2014.01.049

Moriarty, D. J. W., Pollard, P. C., & Hunt, W. G. (1985). Temporal and spatial variation in bacterial production in the water column over a coral reef. *Marine Biology, 85*(3), 285-292. doi:10.1007/bf00393249

Mosher, J. J., Bowman, B., Bernberg, E. L., Shevchenko, O., Kan, J., Korlach, J., & Kaplan, L. A. (2014). Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *Journal of Microbiological Methods, 104*, 59-60. doi:10.1016/j.mimet.2014.06.012

Nacke, H., Thurmer, A., Wollherr, A., Will, C., Hodac, L., Herold, N., . . . Daniel, R. (2011). Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. *PLoS One, 6*(2), e17000. doi:10.1371/journal.pone.0017000

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Nair, P. (2012). Woese and Fox: Life, rearranged. *Proceedings of the National Academy of Sciences, 109*(4), 1019-1021. doi:10.1073/pnas.1120749109

Nakamura, Y., Cochrane, G., & Karsch-Mizrachi, I. (2013). The International Nucleotide Sequence Database Collaboration. *Nucleic acids research, 41*(D1), D21-D24. doi:10.1093/nar/gks1084

Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics, 25*(10), 1335-1337.

Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell, 166*(5), 1103-1116. doi:10.1016/j.cell.2016.08.007

Nguyen, N.-P., Warnow, T., Pop, M., & White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *Npj Biofilms And Microbiomes, 2*, 16004. doi:10.1038/npjbiofilms.2016.4

Nielsen, A. T., Liu, W. T., Filipe, C., Grady, L., Jr., Molin, S., & Stahl, D. A. (1999). Identification of a novel group of bacteria in sludge from a deteriorated biological phosphorus removal reactor. *Applied and Environmental Microbiology, 65*(3), 1251-1258.

Nikolaki, S., & Tsiamis, G. (2013). Microbial diversity in the era of omic technologies. *Biomed Research International, 2013*, 958719. doi:10.1155/2013/958719

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science, 276*(5313), 734-740.

Partensky, F., Hess, W., & Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiology and molecular biology reviews, 63*(1), 106-127.

Pataky, J., Michener, P., Freeman, N., Weinzierl, R., & Teyker, R. (2000). Control of Stewart's wilt in sweet corn with seed treatment insecticides. *Plant disease, 84*(10), 1104-1108.

Pinton, R., Varanini, Z., & Nannipieri, P. (2001). The rhizosphere as a site of biochemical interactions among soil components, plants, and microorganisms.

Pirrung, M., Kennedy, R., Caporaso, J. G., Stombaugh, J., Wendel, D., & Knight, R. (2011). TopiaryExplorer: visualizing large phylogenetic trees with environmental metadata. *Bioinformatics, 27*(21), 3067-3069. doi:10.1093/bioinformatics/btr517

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glockner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research, 35*(21), 7188-7196. doi:10.1093/nar/gkm864

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*(1), 341. doi:10.1186/1471-2164-13-341

Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S. C., Treusch, A. H., Eck, J., & Schleper, C. (2003). Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Molecular microbiology, 50*(2), 563-575.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data

processing and web-based tools. *Nucleic Acids Research, 41*(Database issue), D590-596. doi:10.1093/nar/gks1219

Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., . . . Sloan, W. T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods, 6*(9), 639-641. doi:10.1038/nmeth.1361

Ramette, A., & Tiedje, J. M. (2007). Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial Ecology, 53*(2), 197-207. doi:10.1007/s00248-005-5010-2

Ramette, A., & Tiedje, J. M. (2007). Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proceedings of the National Academy of Sciences, 104*(8), 2761-2766. doi:10.1073/pnas.0610671104

Reeder, J., & Knight, R. (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nature Methods, 7*(9), 668-669. doi:10.1038/nmeth0910-668b

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics, 16*, 133-151. doi:10.1146/annurev-genom-090413-025358

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics, 38*, 525-552. doi:10.1146/annurev.genet.38.072902.091216

Ritari, J., Salojärvi, J., Lahti, L., & de Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated

reference database. *BMC Genomics, 16*(1), 1056. doi:10.1186/s12864-015-2265-y

Rousk, J., Baath, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., . . . Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME Journal, 4*(10), 1340-1351. doi:10.1038/ismej.2010.58

Rousk, J., Bååth, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., . . . Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *The ISME journal, 4*(10), 1340-1351.

Santamaria, M., Fosso, B., Consiglio, A., De Caro, G., Grillo, G., Licciulli, F., . . . Pesole, G. (2012). Reference databases for taxonomic assignment in metagenomics. *Briefings in Bioinformatics, 13*(6), 682-695. doi:10.1093/bib/bbs036

Schloss, P. D. (2009). A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS One, 4*(12), e8230. doi:10.1371/journal.pone.0008230

Schloss, P. D., & Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology, 77*(10), 3219-3226. doi:10.1128/aem.02810-10

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology, 75*(23), 7537-7541. doi:10.1128/aem.01541-09

Schmidt, S., Wittich, R., Erdmann, D., Wilkes, H., Francke, W., & Fortnagel, P. (1992). Biodegradation of diphenyl ether and its monohalogenated derivatives by Sphingomonas sp. strain SS3. *Applied and environmental microbiology, 58*(9), 2744-2750.

Schmidt, V. T., Reveillaud, J., Zettler, E., Mincer, T. J., Murphy, L., & Amaral-Zettler, L. A. (2014). Oligotyping reveals community level habitat selection within the genus Vibrio. *Frontiers in microbiology, 5*.

Seedorf, H., Kittelmann, S., Henderson, G., & Janssen, P. H. (2014). RIM-DB: a taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. *PeerJ, 2*, e494. doi:10.7717/peerj.494

Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., & Gilbert, J. A. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio, 5*(4), e01371-01314. doi:10.1128/mBio.01371-14

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mobile Computing and Communications Review, 5*(1), 3-55. doi:10.1145/584091.584093

Sharon, G., Sampson, T. R., Geschwind, D. H., & Mazmanian, S. K. (2016). The Central Nervous System and the Gut Microbiome. *Cell, 167*(4), 915-932. doi:10.1016/j.cell.2016.10.027

Sharpton, T. J., Riesenfeld, S. J., Kembel, S. W., Ladau, J., O'Dwyer, J. P., Green, J. L., . . . Pollard, K. S. (2011). PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computational Biology, 7*(1), e1001061.

Shrestha, R. K., Lubinsky, B., Bansode, V. B., Moinz, M. B. J., McCormack, G. P., & Travers, S. A. (2014). QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics, 15*. doi:10.1186/1471-2105-15-33

Simpson, E. H. (1949). Measurement of diversity. *Nature*.

Sinclair, L. SEQenv Github Repository. Retrieved from https://github.com/xapple/seqenv

Sinclair, L., Ijaz, U. Z., Jensen, L., Coolen, M. J., Gubry-Rangin, C., Chroňáková, A., . . . Weimann, A. (2016). *Seqenv: linking sequences to environments through text mining* (2167-9843). Retrieved from

Sinclair L, I. U., Jensen L, Coolen MJ, Gubry-Rangin C, Chroňáková A, Oulas A, Pavloudi C, Schnetzer J, Weimann A, Ijaz A, Eiler A, Quince C, Pafilis E. (2016). Seqenv: linking sequences to environments through text mining. *PeerJ Preprints, 4:e2317v1*. doi:https://doi.org/10.7287/peerj.preprints.2317v1

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., . . . Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences, 103*(32), 12115-12120.

Stackebrandt, E., & Goebel, B. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology, 44*(4), 846-849.

Stamatakis, A. (2006). *Phylogenetic models of rate heterogeneity: a high performance computing perspective.* Paper presented at the Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International.

Stefka, A. T., Feehley, T., Tripathi, P., Qiu, J., McCoy, K., Mazmanian, S. K., . . . Nagler, C. R. (2014). Commensal bacteria protect against food allergen sensitization. *Proceedings of the National Academy of Sciences, 111*(36), 13145-13150. doi:10.1073/pnas.1412008111

Stephan, A., Meyer, A. H., & Schmid, B. (2000). Plant diversity affects culturable soil bacteria in experimental grassland communities. *Journal of Ecology, 88*(6), 988-998.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., . . . Bork, P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science, 348*(6237), 1261359. doi:10.1126/science.1261359

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., . . . Waage, J. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome, 4*(1), 62. doi:10.1186/s40168-016-0208-8

Tikhonov, M., Leach, R. W., & Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME Journal, 9*(1), 68-80. doi:10.1038/ismej.2014.117

Topaz, M., Radhakrishnan, K., Blackley, S., Lei, V., Lai, K., & Zhou, L. (2016). Studying Associations Between Heart Failure Self-Management and Rehospitalizations Using Natural Language Processing. *Western Journal of Nursing Research*. doi:10.1177/0193945916668493

Turroni, F., Peano, C., Pass, D. A., Foroni, E., Severgnini, M., Claesson, M. J., . . . Ventura, M. (2012). Diversity of bifidobacteria within the infant gut microbiota. *PLoS One, 7*(5), e36957. doi:10.1371/journal.pone.0036957

Tuszynski, J. (2014). caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.

Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., & Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience, 2*, 16. doi:10.1186/2047-217x-2-16

Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly review of biology, 85*(2), 183-206.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., . . . Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74. doi:10.1126/science.1093857

von Mutius, E. (2016). The microbial environment and its influence on asthma prevention in early life. *Journal of Allergy and Clinical Immunology, 137*(3), 680-689. doi:10.1016/j.jaci.2015.12.1301

Watanabe, K., Kodama, Y., & Harayama, S. (2001). Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of Microbiological Methods, 44*(3), 253-262.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., . . . Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research, 6*.

White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., & Pop, M. (2010). Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics, 11*(1), 152. doi:10.1186/1471-2105-11-152

Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon, 21*(2/3), 213-251. doi:10.2307/1218190

Whitton, B. A. (2012). *Ecology of cyanobacteria II: their diversity in space and time*: Springer Science & Business Media.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews, 51*(2), 221-271.

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology, 6*(2), e1000667. doi:10.1371/journal.pcbi.1000667

Wroblewski, L. E., Peek, R. M., Jr., & Coburn, L. A. (2016). The Role of the Microbiome in Gastrointestinal Cancer. *Gastroenterology Clinics of North America, 45*(3), 543-556. doi:10.1016/j.gtc.2016.04.010

Xia, Z., Bai, E., Wang, Q., Gao, D., Zhou, J., Jiang, P., & Wu, J. (2016). Biogeographic Distribution Patterns of Bacteria in Typical Chinese Forest Soils. *Frontiers in Microbiology, 7*, 1106. doi:10.3389/fmicb.2016.01106

Xiong, J., Sun, H., Peng, F., Zhang, H., Xue, X., Gibbons, S. M., . . . Chu, H. (2014). Characterizing changes in soil bacterial community structure in response to short-term warming. *FEMS Microbiology Ecology, 89*(2), 281-292. doi:10.1111/1574-6941.12289

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution, 39*(3), 306-314.

Ye, D., Siddiqi, M. A., Maccubbin, A. E., Kumar, S., & Sikka, H. C. (1995). Degradation of polynuclear aromatic hydrocarbons by Sphingomonas paucimobilis. *Environmental science & technology, 30*(1), 136-142.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., . . . Glockner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology, 29*(5), 415-420. doi:10.1038/nbt.1823

http://www.nature.com/nbt/journal/v29/n5/abs/nbt.1823.html - supplementary-information

Yuan, C., Lei, J., Cole, J., & Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics, 31*(12), i35-i43.

Zettler, E. R., Mincer, T. J., & Amaral-Zettler, L. A. (2013). Life in the "plastisphere": microbial communities on plastic marine debris. *Environmental Science & Technology, 47*(13), 7137-7146. doi:10.1021/es401288x

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30*(5), 614-620. doi:10.1093/bioinformatics/btt593

Zhang, L., Wang, X., Jiao, Y., Chen, X., Zhou, L., Guo, K., . . . Wu, J. (2013). Biodegradation of 4-chloronitrobenzene by biochemical cooperation between Sphingomonas sp. strain CNB3 and Burkholderia sp. strain CAN6 isolated from activated sludge. *Chemosphere, 91*(9), 1243-1249.