INFERENCE AND INVESTIGATION OF MARINE MICROBIAL COMMUNITY STRUCTURES IN THE GLOBAL OCEANS

by

RANA BASHWIH

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2016

DEDICATION PAGE

I dedicated this thesis to the memory of my father, May his soul rest in perfect peace. Amen. This thesis is also dedicated to my mother May Allah give her long life Amen and my brother and sisters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Marine microbial communities are complex, and represent a serious analytical challenge. The Bayesian model for inference of microbial community structure (BioMiCo) was used to characterize microbial populations using 16S rRNA within polar, tropical, and temperate environmental zones. Global-scale and local analyses were performed on 356 microbial samples and 72853 OTUs within the ICOMM database. Global analysis showed that polar and tropical zones had distinct community structures with high predictive value and little seasonal variation, although seasonal variation was noticeable in the temperate zone. Local analysis on polar communities demonstrated that there were distinct community structures for the Arctic and Antarctic zones. Within the North Atlantic, temporal heterogeneity differed locally, and this impeded the predictive value of models for the entire North Atlantic. Training a model on a single, well-sampled, North Atlantic site, L4 in the English Channel, substantially improved the predictive value of the model. Finally, the model for the L4 site had predictive value for other English Channel sites, but not for more distant sites within the western and eastern North Atlantic. This result appears to be due to differences among North Atlantic sites in the timing of their seasonal community transitions, and because most other sites have not been nearly as well sampled as the L4 site. The only other well-sampled site in the North Atlantic (Bedford Basin) also exhibits regular seasonal transitiona from year to year. Taken together, these results suggest that environmental changes are the primary drivers of marine biogeographic patterns within the North Atlantic.

Four methodological investigations were applied to Arctic and Antarctic samples, and to the samples from L4 station in the English Channel, for the purpose of exploring the impact of how users might choose to make inferences using BioMiCo. The first was an exploration of different ways of defining the predominant OTUs within an assemblage. The size of the assemblage was very sensitive to the method. I recommend defining predominant OTUs as those having >0.01 posterior probability, as this was the most conservative. The second was an exploration of the impact of "burn-in". As expected, increasing burin-in yielded more stable assemblages; however, the burn-in did not need to exceed 1000 iterations. The third was an exploration the effect of training and testing design on prediction of Arctic and Antarctic samples. The results showed that better predictions were obtained from larger training sets of data. However, training on more than 2/3 of the data did not generate significant improvement. Thus, designs such as leave-one-out cross validation can be reserved for cases where the total sample size is very small. Otherwise, uses should run several replicates on data randomly divided into 2/3 training sets and 1/3 test sets. The fourth explored the effect of pre-specifying different numbers of assemblages (the value of L within the model). The results showed that running 25 communities was sufficient. In conclusion, the choices that users make when running the MCMC can impact their results, but, the approach is robust and good results can be obtained with just L=25 if the training data is of a sufficient size, and if a sufficient amount of burn-in is discarded.

LIST OF ABBREVIATIONS AND SYMBOLS USED

OTU	Operational taxonomic unit
3D	3-dimensional
BioMiCo	Bayesian model for inference of microbial community structure
GOS	Global Ocean Sampling Project
ICOMM	International Census of Marine Microbes
L4-DeepSeq	Illumina V6 deep sequence data for English Channel Seawater at station L4
PCR	Polymerase Chain Reaction
QIIME	Stands for Quantitative Insights Into Microbial Ecology (pronounced *chime*)
PP	Posterior probability

ACKNOWLEDGEMENTS

First and foremost I would like to express my sincere gratitude to my supervisor Dr. Joseph Bielawski for his endless guidance and support throughout my graduate years at Dalhousie University. It is my great pleasure to be one of his graduate students. His contributions are sincerely appreciated and gratefully acknowledged. Also, I owe my deepest gratitude to Dr. Katherine Dunn for her assistance and suggestion through out my project. Thank you both for the countless hours you spent helping me through the writing process. A simple acknowledgement is not enough to show my gratitude for all the hard work, patience, and all the support you have given me. Many times throughout my two years I reflected on how lucky I was to work with both of you. I would also like to thank my co-supervisor Dr. Hong Gu and my committee Dr. Robert Lee and Dr. Morgan Langille for their help and advice. Last but not least, I would like to thank my Family. To my mom Thank you for always being there for me, love me, care for me, supporting and believing in me. I can only hope to one day become half the women you are; strong, intelligent, successful and beautiful. I am blessed to be your child. To my brother Rayan and sisters Nouf, Razan and Layan Thank you for your love and support.

CHAPTER 1: INTRODUCTION

1.1 Microbiomics and marine microbial ecology

Microbiomics is defined as the study of the microbiome – the large, uncultivable community of microorganisms that are present in various environments, including the human body (Dove, 2013). Microbes don't occur in isolation and thus they often exist as part of a dynamic pattern of different species populations (Handelsman, 2004). Isolating and sequencing genomes of individual organisms is often inadequate, as they don't represent the full genetic and metabolic potential of the community. To better understand community-level properties, the field of microbiomics has borrowed many concepts and measurement methods from ecology. These help microbiologists to categorize the microbiomes of different habitats, and to compare the microbiomes of different individuals in the presence of different environmental cues. A fundamental property of a microbiome is its biodiversity; the number and distribution of different species existing in a particular habitat. Classification of biodiversity may be made either with an evolutionary perspective (phylogenetic classification) or an ecological perspective (functional classification) (Colwell, 2009). Understanding the drivers of biodiversity is an essential first-step in the study of inter-relationships between co-existing species.

Until recently, the single biggest hurdle to studying microbial biodiversity was that most microbes, in most environments, were uncultivable within the laboratory environment, and thus they were largely unknown to science. The development of next generation high throughput sequencing methods has changed this (Dove, 2013). These techniques make feasible the direct sequencing of the DNA from an environment, thereby circumventing the need for isolation and cultivation of individual microbes. There are

two different strategies for such an "eco-genetic" approach to microbial biodiversity, and they are described in more detail below. This capacity gave birth to the field of microbiomics, and its application has already had major impacts on other biological disciplines, such as terrestrial ecology, marine ecology, and medicine. In the context of ecology, an understanding of the microbial communities and their place in evolution can aid in the prediction of species interactions, their response to environmental changes, and even their co-evolution.

Despite the advancements made in the area of microbiomics, understanding the functional diversity, microbial community structure and the ecological determinants of microbial communities remains a major challenge; this is all the more important given that (Loreau et al., 2001; Prosser et al., 2007; Doney et al., 2012). Carbon and nutrient cycling within the worlds oceans is a clear example; a great diversity of microorganisms occupy all ocean environments, with distinct communities of microbes playing a key role in these important ecosystem processes (Doney et al., 2012). Further, such communities often respond to changes in the environment including climate change, increased temperature, carbon chemistry, nutrient and oxygen content (Doney et al., 2012). However, measurement of their diversity, let alone identifying functionally coherent communities, is a daunting task. Environmental DNA datasets are typically very large and represent a complex pool of a variety of organisms. Obtaining sufficient representation of very rare sequences can be a challenge. Community composition also changes over time, with some species being only conditionally rare (Shade et al., 2013). Such species can be hard to detect at some points in time while at other times they increase in prevalence in response to changes in environmental conditions or community composition (Shade et al., 2013).

Despite these limitations, there is still widespread excitement about the unprecedented opportunity to study microbial ecology on a global scale. One such study was the Global Ocean Sampling (GOS) project (Rusch et al., 2007) which facilitated the study of surface picoplanktonic communities from the Northwest Atlantic to the Eastern Tropical Pacific. This study provided the scientific community with an ocean metagenomic dataset at a truly global scale, and it was used to understand the relationships between gene functional composition and environmental factors (Sunagwa, 2015). Other studies have intensively sampled the marine environment for both microbial and abiotic variables (temperature, seawater chemistry, chlorophyll, etc.) (Fuhrman et al, 2006; Zwirglmaier et al, 2008; Carsey et al., 2011), and such data is helping us understand how microbial biodiversity is distributed in the marine environment. The next phase is to understand how the microbes within a given environment are structured into communities, how the ecology of those communities is related to various environmental drivers, and how the components of those communities (microbial assemblages, or consortia) are related to ecosystem function.

1.2 The non-cultivability of microorganisms, the birth of metagenomics, and the need for taxonomic markers

The traditional approach to isolation and culture of individual organisms has proven either unsuccessful, or at the very least highly inefficient, as a means to bring consortia of microbes into the laboratory to study. The increased knowledge in the field of microbial physiology and genetics that developed from the 1960s to the 1980s, while impressive, was not representative of the entire microbial world. This is evident by the "great plate count anomaly", which showed differences in microbial numbers between

plate dilution and microscopy (Thomas et al., 2010). Thus, the field has long understood there existed a huge gap in our knowledge of microbial ecology. The recent application of high-throughput sequencing technologies to the total DNA extracted from an environment led to a new way of exploring the composition of a microbial niche (Reeder et al., 2010). While a major leap foreword, direct sequencing of total environmental DNA has a major drawback; extraction of the DNA from the environment decouples it from the organism to which it belongs. This led to a new term, the "metagenome", which refers to the total genomic content of a mixed population of microbes (Handelsman, 2004). Community genomics, environmental genomics and population genomics are often used as synonyms for metagenomics.

Although we can never be certain about the organismal origins of all DNA found within a given environment, there are strategies for indirectly inferring the taxonomic composition of a community. The first is to identify selected marker genes within a metagenome using MetaPhlAn to profile microbial communities and compare them to a reference database where those genes have been assigned to known taxonomic categories (Glass et al., 2010). Thus the metagenomic composition of such sequences provides insight into the taxonomic composition of the microbial community from which the metagenome was sampled. The second strategy is to extract total DNA from the environment, but rather than sequencing the entire metagenome, a specific DNA sequence is amplified via PCR and then sequenced (Erlich et al., 1991). This is called amplicon sequencing, and the target DNA sequence is chosen precisely because it is known to be a good taxonomic marker. The hyper-variable regions of the 16S gene are the *de facto* standard taxonomic marker for studies of bacterial community composition.

1.3 Review of 16S as a taxonomic marker for microbial community composition

All Bacteria and Archaea contain the 16S ribosomal RNA (rRNA) gene. The ribosome is comprised of both the large subunit (LSU) and small subunit (SSU) and functions as the site of protein synthesis (an essential function for all living organisms) (Lagesen et al. 2007). While the 16S rRNA gene codes for the SSU in bacteria, both the 23S rRNA and 5S rRNA genes code for the LSU (DeSantis et al., 2006). The 16S rRNA gene is used as a taxonomic marker for identifying bacteria, and for characterizing community composition for several reasons. First, the 16S rRNA gene is relatively short (1.5 kb), and is comprised of a mix of conserved and hyper-variable regions making it efficient to generate 16S rRNA amplicons via PCR. The rapid generation of 16S amplicons via PCR is comparatively faster and cheaper to conventional genomic sequencing techniques thus facilitating the easier phylogenetic classification and characterization of Bacteria and Archaea (DeSantis et al., 2006). Second, since traditional characterization based upon bacterial phenotypic traits can sometimes have limited taxonomic resolution, most taxonomists today consider the analysis of bacterial DNA more reliable than classification based solely on phenotypes (Janda and Abbott, 2007). Third, researchers may, for a number of reasons, want to identify or classify only the bacteria within a given environmental or medical sample (Janda and Abbott, 2007). The reason that the 16S rRNA gene can be employed as a broad taxonomic marker for bacteria is because it has specific regions that have been mostly conserved over time. Parts of its structure have changed very little over evolutionary time (i.e., those parts evolve very "slowly") due to their crucial function in translating mRNA into proteins. But within this gene there are also parts that are more variable than others (i.e., parts that evolve very "fast"). This is due to the structure of the ribosome itself. The way the

ribosomal RNA folds, creating bonds with itself in certain places (conserved regions) while other portions are looped and unbounded (hyper variable regions), determines the rate at which any portion of the gene accumulates variation.

There is the critical need for characterization of species richness and diversity within microbial eukaryotes, and 16S rRNA is not suitable for this. There is a distinct, but homologous gene in eukaryotes, which codes for the 18S rRNA (a component of the 40S small eukaryotic ribosomal subunit), and serves as an alternative marker for identifying and resolving bacteria as separate from plant, animal, fungal, and protist DNA within the same sample (Hugerth et al. 2014; Huse et al. 2008). Homology is the similarity of traits due to shared ancestry, and the structural and functional similarity of the ribosome in prokaryotes and eukaryotes is due to common ancestry of the genes that encode the ribosomal subunits. Specifically, for most Bacteria and Archaea, the main forms of ribosomal RNA settle at the 5S, 16S, and 23S regions of a sedimentation gradient. However, for most eukaryotes, the main forms of ribosomal RNA settle at slightly different regions and thus have different numerical values (e.g., humans have 5S, 5.8S, 18S, and 28S and 40S) (Armougom & Raoult., 2012). The genes encoding the 5.8S and 5S are homologous to the 5S gene of Bacteria and Archaea, the gene encoding the 18S is homologous to the 16S gene, and the gene encoding 28S is homologous to the 23S gene. The three types of rRNA in prokaryotic ribosomes (23S, 16S, and 5S in accordance with their sedimentation rates) have sequence lengths of about 3300, 1550, and 120 nucleotides, respectively (Armougom & Raoult., 2012). Initially, microbial diversity studies involved sequencing the 5S rRNA gene obtained from environmental samples (Armougom & Raoult., 2012). However, the relatively short sequence length of the 5S

gene, and the relatively few phylogenetically informative sites, limits its utility for taxonomic classification purposes.

Unfortunately, primer specificity and coverage have not well been evaluated for 18S. The database of known eukaryotic rRNA is not nearly as well established as for 16S, and polymorphisms have been found in several primer target regions; this suggests that significant levels of eukaryotic diversity may be escaping detection. Thus, there are currently questions regarding the validity of certain primers. Optimization is often required by modifying the primers to cover the most conserved regions (Wang et al., 2014).

The 16S rRNA gene consists of nine hypervariable regions (denoted V1 to V9), which are separated by ten highly-conserved regions (Baker et al., 2003). By connecting these regions with 3D structure and function, the 16S gene has been subdivided into three regions. The Class I region includes the V4, V5, V6, and contains the most important functional parts of 16S rRNA, as it is comprised of the decoding center and the "690 hairpins" (this functional domain is highly conserved in all three phylogenetic domains (e.g., Archaea, Bacteria, Eukaryota)) (Van de Peer et al., 1999; Schuwirth et al, 2005). On the other hand, the Class II region, which includes V3 and V7, is peripheral to the two functional 16S rRNA centers. Lastly, the Class III region includes V2 and V8, which is at the bottom and top respectively of the 3D structure 16S rRNA (Schuwirth et al., 2005). Because the information rich hyper-variable regions can be targeted via primers designed to match the adjacent conserved regions, most studies target only a portion of the 16S gene. While this strategy reduces sequencing cost, it does mean that a specific region must be chosen beforehand, and the correct primers to study bacterial phylogeny must be employed. Recently, Wang et al., (2009) showed that some primer pairs would result in

the uneven amplification of certain species, which leads to an under- or over-estimation of some species within the microbial community. Thus, the choice of region, and the primers, has become an important part of study design, and those studies that make different choices are likely to generate results that are difficult, or impossible, to compare.

The nine hyper variable regions account for the majority of the sequence diversity observed among bacteria species (Van de Peer et al. 1996). Moreover, these regions evolve so quickly that sequence divergence can even occur among the most closely related strains within species (Sogin et al. 2006). This divergence is too great to allow the development of broadly applicable PCR primers. However, because the hyper-variable regions are usually flanked by conserved regions, it is possible to develop so-called bacterial "universal primers" that will amplify segments of the 16S gene that contain hyper-variable regions (Baker et al., 2003). The coverage of such universal primers for 16S rRNA plays a key role in obtaining unbiased estimates of microbial community diversity. When a primer fails to match, or bind to, its target conserved sequence, then the PCR reaction fails and the target 16S will be undetected in that case. This is the reason why 15-20 nucleotides primer sequences must be located in a highly conserved regions for a reliable phylogenetic assessment of the microbial community via 16S (Armougom & Raoult., 2012).

Given that a region has been chosen, and that the sequence data have been obtained, those data must be appropriately processed before inferences about community biodiversity can be reliably made. The data must first undergo quality filtering and de-multiplexing. Quality filtering involves identifying the first quality score below Q30 and truncating the read before that position, and determining if the truncated sequence is at least 75% of the input sequence length, or if it has an ambiguous base call (N characters)

(Eren et al ,2013). Following these steps, each sequence is assigned to an operational taxonomic unit (OTU), which should only be viewed as an approximation to a species-level taxonomic unit. Taxonomic assignment for each OTU is made based on the sequences within a chosen reference database (SILVA: Pruesse et al, 2007; RDB: Cole et al., 2014; Greengenes: DeSantis et al, 2006), and each OTU is attributed as much taxonomic classification as possible. However, depending how sequences are represented within the reference database, it might only be possible to make high-level taxonomic classification for a given OTU (*e.g.*, only to Order). Inferences about biodiversity and community composition are possible only after completion of these tasks.

As very little lateral gene transfer seems to occur between 16S genes, and as their structure contains both highly conserved and variable regions with different evolution rates, the relationships between 16S genes reflect evolutionary relationships between organisms (Armougom & Raoult, 2012). Sequencing the 16S gene is currently the most common approach used in microbial classification as a result of its phylogenetic properties and the large amount of reference 16S gene sequences available for comparison analyses. Clustering of 16S gene sequence according to a cutoff of 97%, or higher, is usually used as the 'gold standard' for identification of groups that approximate the species-level taxonomic unit. Although thresholds are somewhat arbitrary, and hence controversial, a range of 0.5% to 3% sequence divergence is often used to delineate (approximately) the species taxonomic rank (Armougom & Raoult., 2012). As the percent similarity decreases within a "cluster of sequences (i.e., the sequence similarity cutoff increases), more sequences will be clustered into each OTU, leading to a decrease in the total number of OTUs present. For example, Hong et al. (2006) performed successive decreases in cutoffs from 99% to 70% and found an increase in the total

diversity within an OTU, from approximately 10% to 50%. Thus, using a cutoff less than 97%, will increase the chance of clustering different species within a given OTU. Alternatively, some 16S rRNA sequences of >99% have been shown to represent functionally distinct species (Janda and Abbott, 2007). Thus the standard threshold of 97% is merely a convention, which has been adopted because it is best optimized to facilitate the reduction of data volumes while preventing the analysis of sequencing artifacts by clustering them with real sequences. OTU clustering therefore becomes ineffective within this paradigm when the focus is on obtaining an accurate analysis of the true species-level distributions (Patin et al. 2012).

There is an additional problem with the application of 97% cutoff threshold to the same hypervariable region; the approach is very dependent on the composition of the reference databases mentioned above, and this means that consistency in the levels of taxonomic resolution for different OTUs cannot be guaranteed. This is due to the biased distribution of known microbe species identified in the database. Species that are more intensively studied, such as those found in the human gut, will tend to be identified at the species level and well represented within the reference databases. Whereas, less studied microbes will be classified at higher taxonomic levels since the exact species, or closely related species, have not been identified. Moreover, the natural diversity of those less-studies microbes will be less-well-represented within the reference databases. This is a particular problem for marine microbes (which are the focus of this thesis); they tend to be less studied then human microbes making them less likely to have been characterized at the species level.

Despite the above limitations, the inference of community composition via 16S has been extremely effective, and there are now hundreds of studies based on this approach. The literature is far too extensive to review here; however, the study of Gibbons et al. (2013) provides a good illustration of how it can be successfully applied on a global scale. Gibbons et al. (2013) tried to answer the question if bacterial taxa demonstrate clear endemism, like macro-organisms, or does the bacterial community found at any given site recapture the total phylogenetic diversity of the world's oceans. The bacterial sequence data used in this study was obtained from the V6 hyper variable region of the 16S rRNA gene, with all the amplicons being generated using similar procedures as outlined in Huber et al. (2007). Based on this approach Gibbons et al. (2013) found that increasing sequencing depth at a well-characterized reference site (L4-DeepSeq site: Gilbert et al, 2009) yielded greater phylogenetic overlap with the global assessment of diversity within ICoMM (Zinger et al, 2011). By extrapolation, Gibbons et al. (2013) suggested that 1.93×10^{11} sequences from the L4-DeepSeq site would capture all the diversity represented within the world's ocean (as inferred from the ICoMM dataset). Gibbons et al. (2013) suggest that the marine biosphere therefore maintains a previously undetected and persistent "microbial seed bank".

1.4 From 16S amplicons to ecology; measuring and testing microbial biodiversity

Biodiversity is typically measured at the species level, regardless of the organismal group (*e.g.*, algae, lower invertebrates, higher vertebrates, *etc*.). Environmental biodiversity of uncultured microbes differs, however, in that species-level designations are merely operational. The OTU designations depend both on the chosen hyper-variable region and a sequence divergence threshold. This means that

environmentally derived OTUs will reflect different levels of divergence; they could represent ecotypes within species, or collections of metabolically divergent species. The approach has ecological value because of the objectivity of the criteria; *i.e.*, as long as OTUs are defined in the same way, communities can be compared and inferences can be made about community level divergences (Brown et al., 2015).

A fundamental measure of community biodiversity is *Species Richness*, which refers to the number of species present within a particular region under study. Species richness is directly proportional to the biodiversity in that region (Colwell, 2009). Apart from species richness, the *Relative Abundance* of species is also an important measure of biodiversity, and it measures how common or rare a particular species is relative to those in the same region or community. Another community-level metric of biodiversity is *Species Evenness*, which is a measure of how similar the relative abundance of each species is within a given environment. In the context of metagenomics, the notion of a species is simply replaced by the OTU label; I follow this convention and continue to refer to the taxonomic unit as a species when discussing these measures, but in metagenomics these will summarize OTU richness, abundance and evenness.

Species richness and evenness measurements are generally assessed via one of four mathematical indices and these are used routinely in ecology to assess and compare biodiversity. The first is called the Shannon-Weiner diversity index and is represented as:

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

Where, S is the species richness and p_i is the proportion of the i^{th} species in the community. The second index is called the Simpson diversity index (D) and is

$$D = 1 - \sum_{i=1}^{s} p_i^2$$

Like the Shannon index, the Simpson indices are positively related to species richness although the latter is more sensitive to evenness in a community (Colwell, 2009). The third index, called the Berger-Parker index, relies only on evenness in a community and is represented as the reciprocal of "the proportion of individuals in the community that belong to the single most common species, $1/p_{max}$ " (Levin et al,2009). Finally, the fourth index, referred to as the Fisher's α is a geometric progression and can be used as a model for relative abundance (note: there are more alpha diversity measures than the 4 presented above; for a review see Hill et al., 2003). However, the use of this index is hampered by its insensitivity to relative abundance of rare species, and by its complexity (Colwell, 2009). Although each of these provides a measure of biodiversity that can be compared among communities, they reduce the considerable complexity of community level interactions into a single summary statistic. Potentially important structural features of microbial communities are discarded.

Despite the considerable information that is lost in the computation of the broad summary statistics described above, they can serve as the basis for testing community-level divergence, and they will continue to serve as an important analytical tool in microbiomics. Computation of these indices is now automated within programs like EstimateS (Robinson et al., 2010) and QIIME (Caporaso et al., 2010). Further, these

programs allow users to tune the estimators of species richness and diversity to adapt them to each individual scenario. In other words, although the estimators depend on sample size, most of the richness estimates will be stabilized with the sample sizes typically available (Hughes et al, 2001). Several approaches have been developed to compare estimates of species richness from many samples, and one of these is rarefaction (Colwell, 2009). Rarefaction is used to compare richness among sites when the treatments or habitats have been sampled unequally. A rarefied curve is usually obtained by randomly drawing a set of *n* observations from a larger pool of *N* samples multiple times, and plotting the average richness according to the number of individual sampled, or the number of samples taken (Heck et al., 1975). Rarefaction does not address the problem of bias in richness estimators. Bias has only been tested in a few natural communities in which there is a known exact abundance of every species (Colwell et al., 1994; Chazdon et al.,1998).

Given the capacity to assess biodiversity within a sample, the next natural step is to compare the samples to each other. In this setting, the diversity is no longer summarized on a sample-by-sample basis; rather, the distance between pairs of samples is assessed in terms of biodiversity. Ecologists use the term *alpha-diversity* to refer to within-sample species diversity, and *beta-diversity* to refer to differences in species diversity. There are a variety of methods for measuring the distance between samples in terms of biodiversity (Lozupone et al., 2007). Such measurements are summarized as a single matrix of pair-wise distances, and the overall structure of the data can then be visualized as a tree, or a network, or via principle components analysis. Software programs like mothur (Schloss et al., 2009) and QIIME (Caporaso et al., 2010) now

automate the process of making inter-community comparisons, and offer many choices for diversity metrics (Robinson et al., 2010).

The most widely used methods to formally test for differences in microbial community composition among samples include LIBSHUFF (Schloss et al., 2004), TreeClimber (Schloss and Handelsman, 2006), UniFrac (Lozupone and Knight, 2005; Lozupone et al., 2006, 2007) and Analysis of Molecular Variance (AMOVA) (Excoffier et al 1992). LIBSHUFF uses a pairwise distance matrix to compute the probability that two or more communities have the same 16S composition by chance (Schloss et al., 2004). A limitation of the LIBSHUFF-type of analysis is that hypothesis tests must be carried out between all pairs of samples, leading to an increased probability of a false positives without additional multiple test corrections. Moreover, it provides no information about the level of similarity. Like LIBSHIFF, TreeClimber is only to test for community differences (Schloss and Handelsman, 2006). This parsimony-based test requires a user-supplied phylogenetic tree. UniFrac employs a unique metric that summarizes the amount of phylogenetic history that is unique between a pair of samples. The resulting distance matrix is widely used to visualize clusters (*e.g.*, PCoA) or as the basis of formal statistical testing (but requiring multiple test corrections when more than one pair is tested) (Lozupone et al., 2006). The AMOVA test also works on the distance matrix to test for significant variance among samples (as it is inspired by ANOVA, the test is relative to the null hypothesis that all variance arises from sampling from a single, unstructured, community). AMOVA differs from the other three in that it focuses solely on genetic diversity (Martin, 2002). The other three tests can be significant for communities that have the same diversity but different taxonomic composition (but they will also be significant if diversity differs). All these methods can be computationally

costly with big datasets; particularly those that rely on a tree based metric (Lin et al., 2012). Yang et al. (2013) recently developed a novel metric called the compression-based distance (CBD) that quantifies the degree of similarity between microbial communities. Testing based on this technique appears to be considerably faster than other methods while being just as powerful (Yang et al. 2013). The CBD method does not include taxonomic information, so in that way the interpretation of a significant result is more similar to AMOVA.

Although 16S rRNA phylogentic marker sequences are key to the study of microbial community composition, they don't provide direct information about a community's functional capabilities. However, a computational approach called PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved State: (Langille et al., 2013), can be used to predict full metagenome composition from 16S data. This approach uses the complete genomes within a reference database, and an extended ancestral-state reconstruction algorithm, to predict the gene families that are present within a sample, and thereby estimate the metagenome composition. By predicting the abundance of gene families in host-associated and environmental communities, this approach permits inference of community-level function from just the abundance of 16S sequence reads (Langille et al., 2013).

1.5 The BioMiCo framework for modeling microbial community structure

The measures of diversity described above do not capture the potentially complex hierarchical structure that can exist within a single microbial community. Further, microbiome samples often represent mixtures of communities, with each community composed of overlapping assemblages of species. The number of species is usually huge

and abundance information for many species is often sparse. Despite the fact that sequencing of microbial communities is no longer methodologically challenging, the data still pose a significant analytical challenge in capturing the complex hierarchical structure. In addition the sparseness of some species or strains impacts estimates of abundance. For these reasons, classical methods sometimes have limited value for identifying complex features within microbial community data. In order to address this challenge, the Bayesian inference of microbial communities (BioMiCo) model was developed in order to understand how assemblages of OTUs contribute to microbiome structure, and how assemblages are related to the known features of the samples under study (Shafiei et al., 2015).

BioMiCo makes use of abundance data for a given amplicon (typically 16S) obtained from environmental DNA, and models each sample's composition through a two-level hierarchy of mixture distributions that have been constrained by Dirichlet priors. BioMiCo is a Bayesian analytical framework, and uses the prior constraints to overcome the challenges posed by many variables, sparse data and large number of potentially rare species.

BioMiCo is a supervised modeling framework (Shafiei et al., 2015) for building predictive models based on the features of microbial community structure that it resolves within samples. Data analysis is carried out in two separate phases (training and testing), with the predictive community structures (OTU assemblages) resolved from the training datasets. In the training phase the model is applied to part of the data and supplied with labels for the features of interest; this is the phase in which the model learns the assemblages of OTUs to predict the features of interest (Shafiei et al., 2015). In the testing phase, the model is applied to the remainder of the data, but it is not supplied with

the feature labels. In this phase, the model is used to predict the "hidden" labels based on what it "learned" from the training data. The results of the testing phase are used to determine the predictive accuracy of the model (Shafiei et al., 2015).

BioMiCo can be applied to cross-sectional or serially sampled data. Serially sampled data is often considered more valuable for validation of a predictive model, as it allows for easy categorization of "new data" for the testing phase according to time points that follow the training data. Several BioMiCo models have been built for serially sampled datasets and then used to understand and predict transitions between complex communities composed of hundreds of microbial species (Shafiei et al., 2015; El-Swais et al., 2015). Since serially sampled data is not required for supervised methods, many samples can be taken at just a single time point and divided into two parts for training and testing.

The features of interest within BioMicCo are specified with "factor labels", and when the value for a given sample is specified, it's called a "factor value" (Shafiei et al., 2015). It is important to note that generalized factors of interest such as ethnicity, health status or even the identity of the human host can be used as labels in either cross-sectional or serially sampled data. By providing unique indicator values for more than one label, users can also specify multiple factor labels (e.g., summer-deep, summer-shallow, winter-deep, winter-shallow). The analytical objective is then to compute the posterior probability that each test sample originated from a microbiome that had the factor value(s) in which the model was trained on, with discrete assignment to a particular factor value based on the maximum posterior probability (Shafiei et al., 2015).

The structure of BioMiCo is hierarchical. Mixtures of T different OTUs are the basis of L different microbial assemblages (here, assemblages are sets of OTUs that tend

to co-occur in the data) and unique mixtures of assemblages are associated with K different factor values. The values of L (numbers of assemblages) and K (number of factors) are fixed by the user prior to the analysis. The relative contribution of each factor to the n samples is modeled by the probability vector π_n. The contribution of each assemblage to a given factor is modeled by the probability vector θ_k (L mixing probabilities that sum to 1). Thus there are K vectors of L mixing probabilities. The contribution of each OTU to a given assemblage is modeled by the probability vector ϕ_l (T mixing probabilities that sum to 1). Symmetric dirichlet priors are placed on π_n, θ_k and ϕ_l. Gibbs sampling is used to integrate out the latent variables π, θ and ϕ and infer posterior mixture weights (Shafiei et al., 2015).

1.6 Thesis objectives and structure

This thesis is organized around two broad objectives. The first objective is to test the idea that there are ecologically similar microbial communities across widely dispersed sites in the world's oceans (*e.g.*, surface temperate water). The idea is that similar communities assemble because there is little endemism at this level, and the same organisms are subject to similar abiotic factors and biotic interactions at such locations. Testing this is problematic, because the communities are complex, they have complex seasonal transitions, and the summary statistics for alpha and beta diversity shown above do not exploit all the information about community structure. I will employ the BioMiCo framework to address this question because it is focused on assemblages of OTUs. The advantage of this is (*i*) assemblages are more effective at leveraging the structure arising from potentially large numbers of rare OTUs, (*ii*) inferences based on assemblages should be more robust to inter-site variability, and (*iii*) modeling mixtures of assemblages

provides a means of assessing the seasonal transitions between complex communities. The results of this investigation are presented in Chapter 2.

The second objective is to evaluate inferences carried out under alternative analytical strategies that are determined by the user of BioMiCo. Inference about community structure relies on the capacity to reliably estimate the posterior mixture of assemblages within a community, and the mixture of OTUs within an assemblage. The reliability of such inferences could be impacted by how MCMC is used to approximate the posterior probabilities. I will apply alternative methods using the same data from Chapter 2. First, I will investigate different ways to define predominant OTUs. Then I will explore the importance of the burn-in, and the number of MCMC iterations. Next, I will investigate the impact of the design of the training and testing phases on the results. Finally, I will investigate the impact of number of assemblage in the model on the posterior distributions of the predominant OTUs. The results of this investigation are presented in Chapter 3.

I conclude this thesis (Chapter 4) with a review of the main results and I offer some suggestions for directions of future research.

CHAPTER 2: INFERENCE OF MARINE MICROBIAL COMMUNITY STRUCTURES IN THE GLOBAL OCEANS BY USING BAYSIAN METHODS

2.1 INTRODUCTION

2.1.1 The importance of the global ocean microbial community.

Present in every liter of seawater are billions of marine organisms, which represent a structured ecological community that regulates biogeochemical processes. It is not hyperbolae to say that the marine community affects the entire earth biome (*e.g.*, autotrophic energy conversion/production). Marine based microbes are recognized as drivers of major biogeochemical processes such as photosynthesis and cycling of nitrogen, phosphorus and other nutrients (Sunagawa et al., 2015). The collective metabolism of microbial communities in marine environments have been shown to have global effects on fluxes of energy and matter in the sea, on the composition of the earth's atmosphere, and on global climate (Sunagawa et al., 2015). The functioning of these communities also determines how the global ocean will respond to environmental changes, both natural and anthropogenic (Hansen et al,2001;Doney et al., 2012;Sunagawa et al., 2015). However, understanding microbial community structure (both globally and locally), functional diversity, and their ecological determinants remains a challenge because some communities are not stable over time or location (Fuhrman et al 2015; Sunagawa et al., 2015).

As communities, microbes interact with each other and respond collectively to disturbances in their surroundings (Sunagawa et al., 2015). As the environment is altered, the microbial population shifts and changes, but changes to the microbial population may cause feedback changes to the (local) environment. For example, Sunagawa et al. (2015)

found that, of the environmental parameters that influence the formation of microbial communities within sunlit oceans, temperature was identified as the major predictor of changes in community composition. Thus, one implication is that global warming has, and will continue to have, a causal role in changes to microbial communities that accrue over time. Through the combined activity of changing microbial communities the ocean's chemistry, and even measures of the "habitability" of the entire planet, could change as well (Sunagawa et al., 2015).

2.1.2 The global ocean seed bank hypothesis

Simplistically, the global seed bank hypothesis (Gibbons et al., 2013) argues that every part of the marine environment contains all the microbial species found in all parts of the marine environment. What changes from location to location is the relative proportion of each species. The relative abundance is determined by local environmental factors (e.g., temperature, light, nutrient density and specificity), the interactions of microbes (e.g., competition), and the presence of predators.

The global seed bank hypothesis is a specific instance of Baas Becking's (1934) hypothesis that everything is everywhere but variation in environmental factors drives biogeographic patterns of microbial communities, which are altered by biological interactions (Gilbert et al., 2009). As niches open and close, community composition changes across space and time; these changes would be facilitated either by rapid dispersal, or by rapid growth of rare or dominant taxa from a "microbial seed bank".

After deep sequencing of individual microbial communities from the L4-DeepSeq datasets using approximately 10 million 16S rRNA V6 reads (to identify bacterial OTUs), Gibbons et al. (2013) demonstrated that almost all OTUs that were previously identified

at any time during a 72-month times-series from the Western English Channel site were detectable in the single L4-DeepSeq time point. The authors concluded that, in this ecosystem, all the taxa were present at all times but their relative abundance varied over many orders of magnitude as environmental conditions changed. In contrast to the widely accepted model that the presence or absence of particular microbial taxa drives community structure (Caporaso et al., 2012), the above result suggests that global patterns of bacterial community composition consists primarily of changes in relative abundance of shared community members.

The hypothesis that suggests that all bacteria are found in any particular environments as a result of an immense and persistent microbial seed bank can be tested. The alternative is that certain environments lack some bacteria. In order to test this hypothesis, Gibbons et al. (2013) compared L4-DeepSeq dataset with about 356 datasets of bacterial 16S rRNA V6 amplicon sequences from over 40 site of the global International census of Marine Microbes (ICoMM) sample locations. These data are comprised of different studies which range from marine pelagic and sediment samples to mangrove and sponge associated environments. Such comprehensive sampling allowed researchers to investigate the overlap in community membership between biomes, phylogenetic similarity of communities in different biomes, and the probability that core global ocean microbiota might be identified from deep-sequencing.

2.1.3 Are there characteristic community structures at the global level (polar, temperate, and tropical), or are local differences the major determinates of structure?

The ocean's surface has been divided into four main temperature zones and include; the tropical, the warm-temperate, the cold-temperate and the polar zones which has further been sub-divided by ocean basins and adjacent landmasses. Marine microbes have evolved in response to various features of those environments, with temperature presumed to be a major factor that regulates species distributions (Nishiguchi, 2000). However, factors such as light level, which is affected by depth of water, sediment levels and salinity are also important, and these represent a significant source of variability on local scales (Nishiguchi, 2000).

Ghiglione et al (2012) compared bacterioplankton diversity between polar oceans by pyrosequencing the V6 region of the small subunit ribosomal (SSU) rRNA gene. After comparing Antarctic and Arctic Ocean surface communities, they found that over 70% of OTUs were unique to Arctic Ocean and over 78% were unique to the Antarctic Ocean. There was more dissimilarity between coastal surface Arctic and Antarctic Ocean communities, than in their respective open ocean communities, while the deep ocean communities differed even less. This study, therefore, suggests that difference in environmental conditions at the poles create different selection mechanisms controlling surface and deep ocean community diversity. If there is a global seed bank, those differences may also be causing the assembly of unique community structures. Marine microbe communities are not necessarily stable over time. This would be most obvious for temperate environments, which have the greatest seasonality. Diverse and temporally dynamic bacterioplankton communities often inhabit temperate oceans, but

understanding how those communities change at different time scales remains a challenge. El-Swais et al (2015) combined time series observations with molecular analysis of formalin-fixed samples from a coastal inlet of the northwest Atlantic Ocean. They showed that the combination of temperature, nitrate, small phytoplankton and *Synechococcus* abundances were best predictors for roughly 38% of variability in annual bacterioplankton communities. They found that over 32% of community variability over spring bloom development could be explained by silicate, while 16-27% of community variability during the transition into and out of the autumn bloom could be explained by nanophytoplankton and picophytoplankton levels. El-Swais et al. (2015) employed the supervised BioMiCo framework to identify local microbial assemblages characteristic of different seasons (winter, spring summer & fall), and resolved how local community transitions occur between the seasons, as well as around seasonal algal blooms.

This chapter has three broad objectives. The first is to characterize community assemblages on a global scale. This determination of "global structure" in marine communities is at the broadest level of analysis (polar, tropical and temperate zones). The second objective is to understand the extent to which that structure has predictive validity or utility, and to investigate to what extent local community dynamics (*e.g.*, temporal community transitions) might negatively impact predictive value. The third objective is to investigate community transitions over a seasonal cycle within the temperate zone. Five datasets are used to meet these objectives. The datasets were assembled by including or excluding specific sampling locations of the full-ICoMM dataset, and adding other relevant public datasets where available. Differences among these datasets reflect the need to progressively narrow the scope of the investigation: (*i*) global, (*ii*) polar, (*iii*) North Atlantic, (*iv*) a single sampling station (L4) and (*v*) English Channel.

2.2 METHODS

The research questions of this thesis are addressed by using five datasets comprised of the 16S rRNA V6 hypervariable region sequenced from marine environmental DNA. The first is comprised of the global International Census of Marine Microbes (ICoMM), which is comprised of samples from around the world (Gibbons et al., 2013). Hereafter, this dataset is referred to as the "full-ICoMM dataset". The second dataset is restricted to the polar samples from full-ICoMM dataset, and is combined with polar ocean samples from other studies and locations. This dataset is referred to as "Polar dataset", and it is used to investigate differences between northern and southern polar oceans. The third dataset is referred to as the "North Atlantic dataset", and it is comprised of all North Atlantic (temperate zone) samples from full-ICoMM data. The fourth dataset is referred to as "L4 dataset" and it's comprised of the samples from a single station (L4) within the full-ICOMM data. The fifth dataset was created by combining the L4 samples with other samples from the North Atlantic. This dataset is referred to as "expanded English channel dataset", and it is used to investigate seasonal transitions between assemblages within the English Channel.

2.2.1 Full-ICoMM dataset

The ICoMM dataset is comprised of 356 bacterial samples from different locations around the world (Zinger et al, 2011). These samples catalogue 72854 different microbial OTUs (defined according to 97 % sequence similarity threshold for 16S) based on sequences of the 16S rRNA V6 hypervariable regions (Zinger et al, 2011). This rich dataset offers the opportunity to understand microbial biodiversity on a global scale. For

each sample the following additional information was available: absolute depth, sample source as classified by International Hydrological Organization (IHO classification: 26 categories), environmental biome (20 categories), environmental features (27 categories), environmental material (19 categories), environmental habitat (7 categories), latitude, longitude, sample type (4 categories), and sample season. The categories for environmental biome, feature, material, habitat and sample type are given in Appendix Tables A1 to A6.

2.2.2 Polar dataset

The 39 polar samples from the ICoMM dataset were analyzed for differences between the northern (Arctic) and southern (Antarctic) polar oceans. These samples represent a subset of those used in the previous analyses, and were comprised of 25 Arctic and 14 Antarctic samples. Hence, the sample characteristics are quite similar.

The polar samples from the ICoMM dataset were then combined with additional samples collected from both Arctic and Antarctic sources (Deep Arctic Ocean (Galand et al., 2010), Arctic Chuki and Beaufort Sea (Cottrell & Kirchman.,2009), Palmer Station (Luria et al., 2014), the Amundsen Sea Antarctic (Ghiglione et al., 2012), and the Census Antarctic Marin (Murray & Grzymski.,(2007).). Combining these data with the polar ICoMM data produced a dataset containing a total of 70 samples, with 24 from the Arctic and 46 from the Antarctic. Collectively those samples contained 13313 microbial OTUs. The additional data samples were processed using Dr. Langille's standard operating procedure for 16S data (https://github.com/mlangill/microbiome_helper/wiki/16S-standard-operating-procedure).

2.2.3 North Atlantic dataset

This dataset is comprised of the 124 North Atlantic *water column* samples from the ICoMM dataset. Only those samples from the temperate zone were retained for this dataset. This subset of samples contained 16384 microbial OTUs. These samples were taken from different depths; ranging from 0 to 5034 meters. It is important to note that the samples were taken during different seasons, and that they were sorted and grouped by season. For this analysis, there were 13 winter samples, 53 spring samples, 21 fall samples and 37 summer samples. Of the 124 samples, 95 were from coastal waters (fall: 17, spring: 41, summer: 24, winter: 13), 17 from open oceans (fall: 4, spring: 8, summer: 5, winter: 0) and 12 from vents (fall: 0, spring: 4, summer: 8, winter: 0). Additional features associated with the samples were: marine wind mixed layer (79); near an island (16); marine bulk water (1); mesoscale marine eddy (13); or sandy beach (3) areas.

2.2.4 The L4 dataset and the expanded English Channel dataset

The L4 dataset contains 68 L4 samples from full-ICoMM (12 winter samples, 20 summer samples, 20 spring samples, 16 fall samples).

The L4 dataset was expanded by including additional English Channel samples from VAMPS (The Visualization and Analysis of Microbial Population Structures) database. The total number of L4 with the additional English Channel samples was 83 (18 winter samples, 22 spring samples, 28 summer samples, 15 fall samples). These samples were then combined with samples from another site. Some additional samples were derived from the Helogland Roads dataset (Lucas et al., 2015), which is near the English Channel (16 samples). The total number of samples in the expanded dataset was 99 samples.

2.2.5 Analysis of community diversity and structure

The average alpha diversity for each zone, region, or site was computed using four measures (see Chapter 1): Richness (species count), Shannon's index, Simpson's index, Chao1 (a modification to accommodate the fact that rare species may not get into a sample) (Engel et al, 2013). Biodiversity was computed for all 356 samples using the program QIIME (Caporaso et al., 2010) and, where necessary, it was averaged to obtain a summary of biodiversity in each zone. Beta diversity, the diversity between *different* samples, was used to assess the differences between zones or seasons. Beta diversity, as measured by unweighted UniFrac (Lozupone et al., 2007), as implemented within the QIIME program. Unweighted UniFrac is qualitative measure of the difference between communities according to phylogenetic composition (i.e., presence/absence within the data). The BioMiCo framework was used to characterize community assemblages with respect to the following features of interest: (i) polar, temperate and tropical zones, (ii) Arctic versus Antarctic regions, and (iii) seasonal states (winter, spring, summer and fall). Analyses were divided into training and testing phases for all datasets, although the design of those phases differed among the datasets (described below). During training, BioMiCo combines the information across the training samples to learn the assemblage of OTUs that are associated with the features of interest listed above. The testing phase was then used to evaluate the accuracy of predicting those labels according to their characteristic assemblage structures.

2.3 RESULTS and DISCUSSION

2.3.1 Global-scale investigation: polar and tropical zones have distinct community structures with little seasonal variation

The full-ICoMM dataset consisted of 356 samples. BioMiCo was trained to classify and predict the zone as polar, temperate or tropical using 60% of the samples from each zone within the full-ICOMM dataset. Thereafter, it was tested using the remaining 40%. The trained model was assessed according to the degree to which it could properly predict the identity of the test samples on the training dimensions (polar, tropical, temperate). Five different analytical replicates were run based on random assignment of 60% of the samples to the training set (60% from each environmental zone), and placing the remaining 40% in the test set. Each analytical replicate used a burn-in of 5,000, and was ran on 10,000 iterations. The results were aggregated across all 5 runs even though the test cases are not completely independent. The cases are not completely independent because some samples will be used in more than one training set, and some in more than one test set. Overall performance was based on a total of 715 test cases derived from the 5 analytical replicates (143 test samples per analytical replicate).

Results summed over the 5 replicates are shown in Table 2.1. There are some important observations. Firstly, BioMiCo is reasonably successful, having achieved 72 % correct classification of the test data. This is evident in Table 2.1A, as most of the counts lie along the main diagonal (521 of 712 test cases yields 72 %). All of the polar samples, and most of the tropical samples, were correctly classified. However, the temperate samples were misclassified more than 1/3 of the time (176 of 507, or 35%).

Classification of each test sample was based on the maximum of the posterior probabilities that a sample was from each of the three zones, and these are illustrated in Figure 2.1A. This plot clearly illustrates that the temperate zone samples are poorly classified because there are many of those samples had a high posterior probability for the polar label. The overall accuracy for the temperate samples was about 65%. Because there are so many more temperate samples than polar and tropical (combined), this error makes the trained model seems less successful than it is.

These data do indeed contain strong signal about microbial assemblages. Figure 2.1B shows that the trained model resolved three distinct assemblages of OTUs; one for each zone (Appendix B provides the separate results for each run). This result is not inevitable; given that the model was permitted to use up to 25 assemblages to explain the data (i.e., L=25 in the model), the trained model could have used mixture of several assembles to explain the structure of each zones. Given that the model was very accurate in predicting the polar and tropical zones (Figure 2.1A), and the posterior distribution of assemblages was so concentrated (Figure 2.1B), I conclude that there is strong signal for distinct assemblage structures in the polar and tropical zones. Note that both zones have high posterior probabilities for a single assemblage in the training data.

The difficulty in identifying the temperate samples appears to arise from that there were temperate samples that did not have high posterior probabilities for any of the three assemblages resolved in the training data (Figure 2.1A). Collectively, these results suggest that temperate samples are more diverse, and more dynamic, than either polar or tropical samples. This is supported by basic measurements of alpha and beta diversity. Several measures of alpha diversity are shown in Table 2.2A, and the temperate zone has the highest averages for all measures. Within-zone beta diversity unweight UniFrac

(Table 2.2B) was highest for the temperate zone, indicating greater among-sample community variability within that zone.

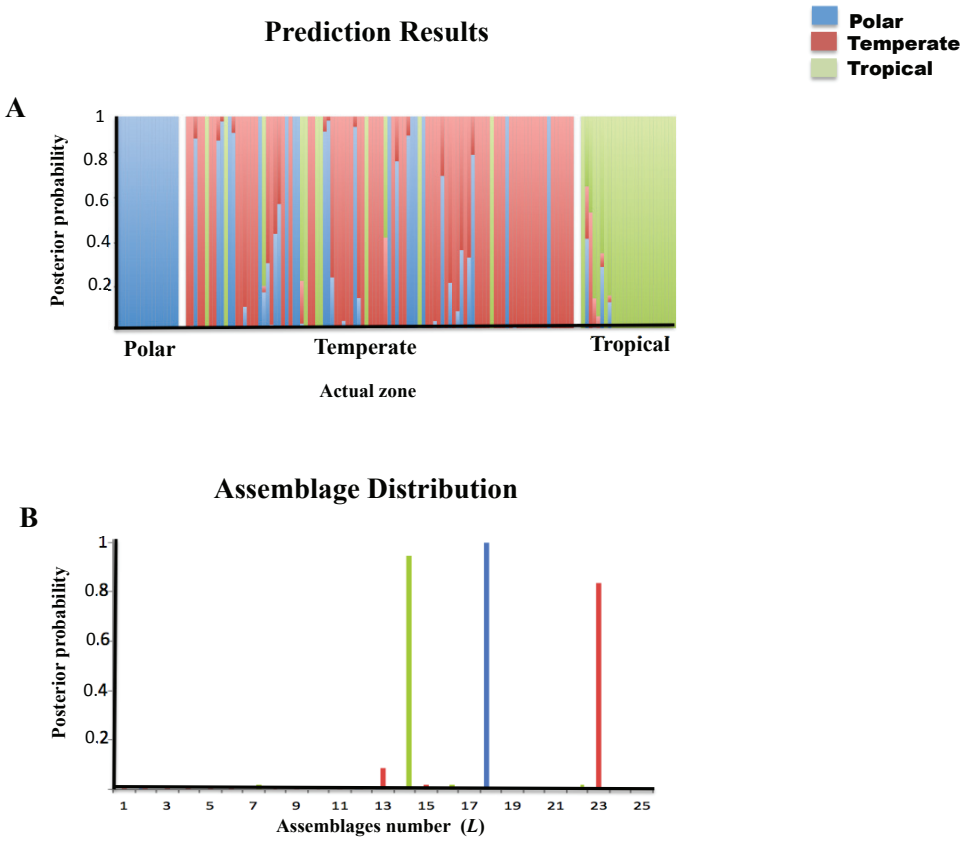

Figure 2.1: Bayesian prediction of polar, temperate and tropical community structures. (A) Prediction results from BioMiCo for the three zones. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. The posterior predictions are derived from just 1 of 5 cross-validation replicates (other were similar). (B) The posterior distribution of OTU assemblages. Note that three assemblages are responsible for the vast majority of posterior probability despite allowing the model to use as many as 25 assemblages to explain the data.

Table 2.1
Prediction results summed over 5 replicates of cross-validation. (A) Counts of the correct and incorrect classifications of the tests samples for the three zones. (B) Classification accuracy shown as a percentage. Note that for both (A) and (B) the results for the correct classification are given along the diagonal.

A

Prediction results

		Actual			
		Polar	Temperate	Tropical	Total
Predicted	Polar	80	124	10	214
	Temperate	0	331	5	336
	Tropical	0	52	110	162
	Total	80	507	125	712

B

prediction as a percentage

		Actual		
		Polar	Temperate	Tropical
Predicted	Polar	100.00	24.70	8.00
	Temperate	0.00	65.00	4.00
	Tropical	0.00	10.30	88.00
	Total	100.00	100.00	100.00

Recall that each run used a different random selection of samples, and that each sample would contain a different set of OTUs. Although classification is based on complex co-occurrence patterns among OTUs, in which large numbers of rare OTUs can be informative (Dunn et al., 2016), the patterns of concordance (or discordance) among separate analytical runs can be used to further investigate model performance. Here, predominant OTUs within an assemblage are defined as having a posterior probability (PP) >0.01, and they might be expected to exist in "most" samples, and as such, would be expected to become an important part of an assemblage for a given run. Table 2.3 shows the number of runs in which each OTU (with a PP $\geq$ 0.01) was identified for each of the three assemblages (polar, temperate and tropical). For example, OTU 5901 was inferred to be "predominant" in 1 of 5 temperate assemblages and in 5 of 5 tropical assemblages. The counts ranged from 0 in all runs (not shown for clarity) to 5 in all runs. Table 2.3 summarizes the counts of predominant OTUs within the three major assemblages (polar: 20; temperate: 8; and tropical: 14) over the 5 separate runs. The important finding here is that the distribution of predominant OTUs is much sparser in the temperate assemblage, and shows considerably more among-run variability. This is expected if the assemblage structure of the temperate zone is more dynamic over time.

Table 2.2
Measures of alpha (A) and beta (B) diversity for the polar, temperate and tropical zones. Beta diversity was measured using the unweighted UniFrac .

A **Alpha Diversity**

	Polar	Temperate	Tropical
N	39	255	62
Richness	363.4	667.0	611.3
Shannon's Index	5.58	7.03	6.67
Simpson's Index	0.908	0.963	0.934
Chao 1	540.5	1158.1	1039.2

B **Beta Diversity**

Zone	Polar	Temperate	Tropical
N	39	255	62
Beta-diversity	0.361	0.440	0.426

Table 2.3.
The predominant OTUs (posterior probability > 0.01 criterion) for 3 zones.

OTU ID	Polar	Temperate	Tropical	Number of Zones
5901		1	5	2
8407	5			1
84240	2	5		2
95477	2			1
105551		1		1
105774	4			1
151578	5			1
158847			2	1
234682		1	3	2
277633			5	1
306657	5			1
317182		4		1
317708	4			1
319540			4	1
511577			4	1
528099	1			1
534609		3		1
543795	1			1
557211			2	1
New.0.Reference.OTU103333	3			1
New.0.Reference.OTU105656			5	1
New.0.Reference.OTU106808	4			1
New.0.Reference.OTU62958	1			1
New.0.Reference.OTU72937	5	5	5	3
New.0.Reference.OTU91013	5			1
New.0.Reference.OTU377	2			1
New.0.Reference.OTU398	2			1
New.1.Reference.OTU1050			1	1
New.1.Reference.OTU1479	1			1
New.1.Reference.OTU1750	1			1
New.1.Reference.OTU1790			1	1
New.1.Reference.OTU1798			1	1
New.1.Reference.OTU2129		1		1
New.1.Reference.OTU322	5			1
New.1.Reference.OTU390			1	1
New.1.Reference.OTU592			1	1
New.1.Reference.OTU769	1			1
Column Total	20	8	14	37

Note: The taxonomy for each OTU ID in Appendix (C)

Based on these results, I hypothesized that temporal dynamics associated with seasons might have impacted the findings. Therefore, as a means to improve prediction, I divided the temperate zone into four seasons using the collection date of each sample and the hemisphere (northern or southern hemisphere) in order to split temperate to seasons -- fall, spring, summer and winter. BioMiCo was used to predict 6 environmental labels: polar, tropical, temperate-summer, temperate-spring, temperate-fall, temperate-winter. Again the training was on 60% of each label, and testing was on the remaining 40%. Each run had 145 test cases. Across all 5 runs there were a total of 725 test cases. The average results for all five runs are shown in Table 2.4. Surprisingly, the overall success at classification of the test samples dropped to 37.7% (recall that with just three labels, the overall success was 72.9%). However, the trained model was still successful at predicting the polar and temperate-winter samples according to community structure. That is, 71 of 80 polar samples and 44 of 45 winter samples were correctly classified. The remaining samples are poorly predicted most of the time.

It is interesting that the assemblage distributions shows good signal for each of the 6 environmental labels (Figure 2.2B). This means the training phase identified six highly informative assemblages, with each one of these assemblages contributing a high posterior probability to one of the four seasons. These results seems to indicate that the samples from different seasons were clearly distinguishable from each other. However, looking at the test data, the problem appears to be that many temperate (winter, spring, summer and fall) samples are best described as a mixture of 2 or more assemblages (47 temperate samples; see also Figure 2.2A). Furthermore, there also are cases when the classification is unambiguously wrong (14 temperate samples). Also, it's noticeable that

the accuracy of predicting tropical zone samples decreased after adding the season label to the temperate samples (see Figure 2.1, also see Figure 2.2). With the exception of species richness, alpha diversity is consistently higher during the spring and fall seasons within the temperate zone. Within-zone beta diversity was lowest for polar and temperate-winter zones (Table 2.5B), which also happened to have good classification rates in the test data. Within-zone beta diversity for the temperate samples was comparable to, or higher than, the tropical zone, despite the temperate zone being subdivided. Thus, if the poor classification results are indeed due to temporal heterogeneity, then division of the temperate zone samples into broad seasons (at least at the global level) was not sufficient to capture enough variability to permit reliable predictions.

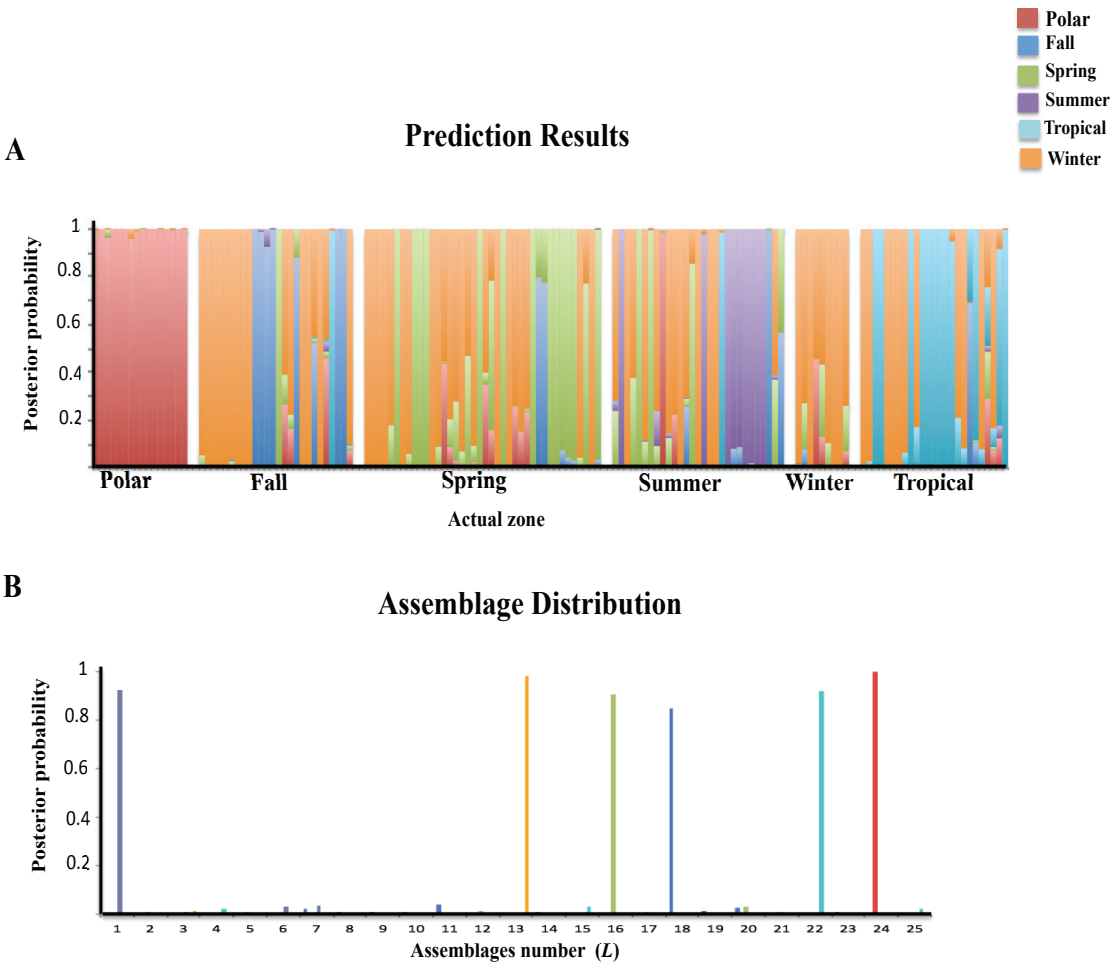

Figure 2.2: Bayesian prediction of polar, tropical and temperate community structures divided into four seasons (winter, spring, summer and fall). (A) Prediction results from for the six possible sample categories. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. The posterior predictions are derived from just 1 of 5 cross-validation replicates (other were similar). (B) The posterior distribution of OTU assemblages. Note that six assemblages are responsible for the vast majority of posterior probability despite allowing the model to use as many as 25 assembles to explain the data.

Table 2.4
Prediction results summed over 5 replicates of cross-validation. (A) Counts of the correct and incorrect classifications of the test samples for the six sample categories. (B) Classification accuracy shown as a percentage. Note that for both (A) and (B) results for the correct classification are given along the diagonal.

A

Prediction results

		Actual						
		Polar	Winter	Spring	Fall	Summer	Tropical	Total
Predicted	Polar	**71**	0	3	1	1	2	78
	Winter	9	**44**	144	94	89	71	451
	Spring	0	0	**44**	4	9	0	57
	Fall	0	0	4	**26**	3	3	36
	Summer	0	0	2	2	**40**	1	45
	Tropical	0	1	3	3	3	**48**	58
	Total	80	45	200	130	145	125	725

B

prediction as a percentage

		Actual					
		Polar	Winter	Spring	Fall	Summer	Tropical
Predicted	Polar	**88.75**	0.00	1.50	0.77	0.69	1.60
	Winter	11.25	**97.78**	72.00	72.31	61.38	56.80
	Spring	0.00	0.00	**22.00**	3.08	6.21	0.00
	Fall	0.00	0.00	2.00	**20.00**	2.07	2.40
	Summer	0.00	0.00	1.00	1.54	**27.59**	0.80
	Tropical	0.00	2.22	1.50	2.31	2.07	**38.40**
	Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 2.5
Measures of alpha (A) and beta (B) diversity for six possible sample categories: polar, tropical, temperate-winter, temperate-spring, temperate-summer, and temperate-fall.

A **Alpha Diversity**

	Polar	Winter	Spring	Summer	Fall	Tropical
n	39	21	98	72	64	62
Richness	363.4	589.2	591.7	588.7	895.9	611.3
Shannon's Index	5.58	6.92	6.80	6.81	7.65	6.67
Simpson's Index	0.908	0.969	0.959	0.962	0.970	0.934
Chao 1	540.5	970.4	1039.7	958.1	1625.9	1039.2

B **Beta Diversity**

Zone		Polar	Winter	Spring	Summer	Fall	Tropical
	N	39	21	98	72	64	62
Bate-diversity		0.361	0.377	0.431	0.452	0.412	0.426

This investigation showed that both polar and tropical zones seem to have distinct community structures. In contrast, the temperate zone did not exhibit a distinct community structure; rather its taxonomic composition broadly overlapped the other zones and had more taxonomic variation among samples. The structures within the temperate samples varied greatly among seasons. However, the predictive value was not great. There may be other factors that need to be considered. For example, the ICoMM data is global. The temperate samples may represent greater environmental heterogeneity as a function of global location (when compared to the polar or tropical zones). Seasonality is more pronounced for the temperate zone and this seasonality (the timing of the seasons) may occur at different times. This effect would interact with the fact that the samples were collected under a wide variety of conditions (see Appendices A1-A6). All of these factors would make it difficult to construct representative assemblages for predictions. Also, temperate samples share many predominant OTUs with polar and tropical (see table 2.3) but they have characteristic differences in their relative abundance.

2.3.2 Polar investigation: the Arctic and Antarctic zones have distinct communities

The polar ocean investigation was carried out in two phases. The first phase was based on just the polar samples from the full-ICoMM dataset, which contained 39 such samples (25 Arctic and 14 Antarctic). All these samples were taken from the polar ocean water column. Leave-one-out cross-validation was applied to this dataset. In the leave-one-out strategy, one sample is removed and the remaining 38 samples are used for training. The trained model is then used to predict the one sample left out (this one sample is referred to as the 'test sample' in this setting). Thus, the procedure was

repeated for each of the 39 samples. A1500 burn-in period with 20,000 iterations was used for each repetition.

Prediction for the Arctic and Antarctic samples was much improved as compared to the global-scale analyses (see section 2.3.1) that included the ICoMM temperate samples (Figure 2.3A). Classification accuracy was 85.7% for the Antarctic samples and 96% for the Arctic samples (Table 2.6). Interestingly the Arctic samples had consistently higher measures of alpha diversity (Table 2.7), indicating that reliable predictions can be made for diverse communities. Indeed, Shannon's diversity and Simpsons' diversity are comparable to the values inferred for the winter-, spring-, and summer-temperate samples in the previous investigation (Table 2.5A). The posterior distribution put all the weight on just two assemblages, one for the Arctic and one for the Antarctic (Figure 2.3B). These results indicate that there is strong signal for two distinct polar communities.

To investigate the robustness of the results for the Arctic and Antarctic, the polar samples (from the ICoMM data) were combined with additional water column samples from the Arctic and Antarctic (see methods for details). Because this second phase of the analysis was based on a larger dataset (70 samples: 24 Arctic and 46 Antarctic), it was possible to use the 2/3 training and 1/3 testing cross-validation design, as compared to the leave-one-out analysis. Cross-validation was run on 10 random replicates, with 1500 burn-in and 10,000 iterations used for each replicate. The overall prediction accuracy for all the 10 replicates 96.5% for the Antarctic samples and 79.7%for the Arctic samples (Table 2.8). In order to investigate the effect of an even longer run, the analysis was re-run using a 1500 burn-in and 20,000 iterations, and very consistent results were obtained (compare Figure 2.4A and B).

Taken together, the results of both phases of this analysis indicate that the polar oceans have distinct microbial communities. Figure 2.5 shows that the Arctic and Antarctic share many predominant OTUs, but they have characteristic differences in their relative abundance. Note that BioMiCo uses all the OTUs (not just the predominate ones) to determine the posterior probability of a factor label such as Artic or Antarctic (Appendix D gives a more complete summary of posterior distribution of OTUs for the Arctic and Antarctic assemblages, and also illustrates substantial differences in OTU co-occurrence patterns). Examination of the taxonomy using the >0.01 posterior probability for Artic and Antarctic showed that only limited taxonomic resolution was possible for the predominant OTUs; some of them could be identified to the family, and many just to the level of order or class. Noticeable differences in relative abundance were observed for OTUs from Haptophyceae, Oceanospirillales, Flavobacteriales (greater prevalence in Antarctic) and the SAR324 and ZA3648 groups of OTUs (greater prevalence in Arctic). The SAR324, which represents a group of Deltaproteobacteria, is interesting, as it is characterized by extensive metabolic versatility (e.g., Sheik et al. 2014). Members of the SAR324 group are among the most frequently encountered 16S-derived marine OTUs, which is why their metabolic capabilities have been so well characterized. Unfortunately, such information is not yet available for the other predominant OTUs uncovered here.

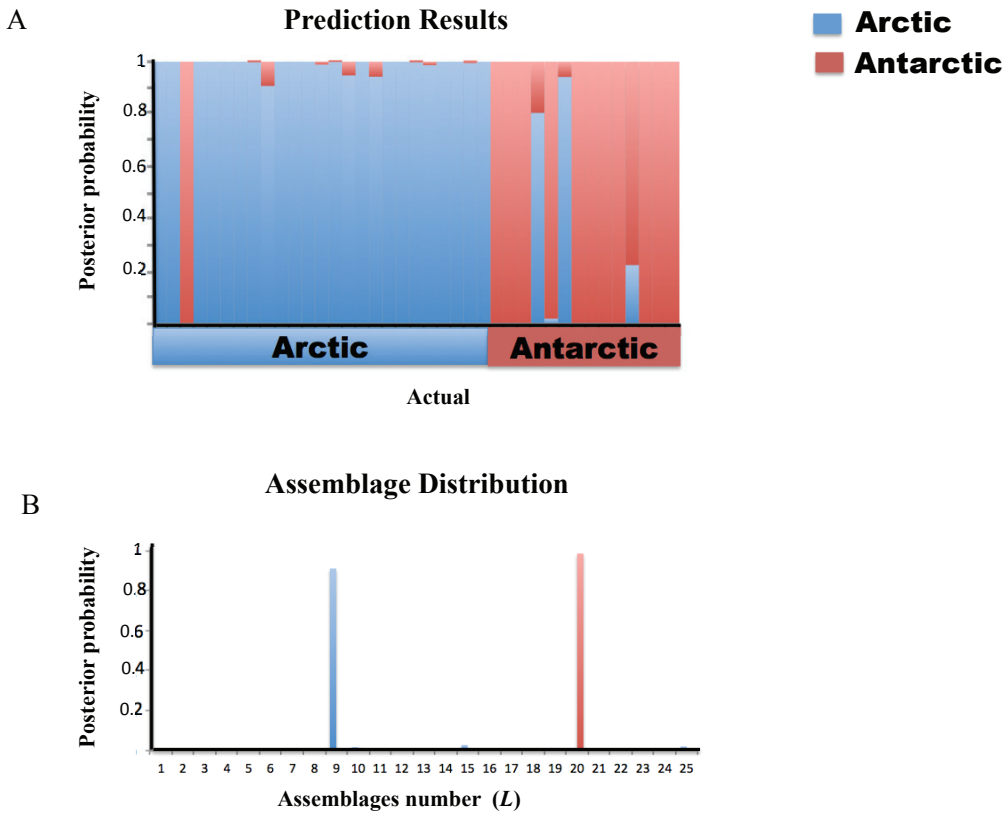

Figure 2.3
Bayesian prediction for polar (Arctic and Antarctic) zones. (A) Prediction results for the Arctic and Antarctic samples. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. Samples are plotted along the x-axis; for clarity the Arctic and Antarctic samples are grouped. The posterior predictions are derived from the leave-one-out strategy for cross-validation. (B) The posterior distribution of OTU assemblages. Note that two assemblages are responsible for the vast majority of posterior probability despite allowing the model to use as many as 25 assembles to explain the data.

Table 2.6.
Predictions for Arctic and Antarctic samples

		Actual		
		Arctic	Antarctic	Overall
Predicted	Arctic	24(96%)	2(14.3%)	26
	Antarctic	1 (4%)	12(85.7%)	13
	Overall	25	14	39

Table 2.7.
Alpha diversity for Arctic and Antarctic samples

	Arctic	Antarctic	Difference
Richness	433.8	170.0	263.750
Shannon's	6.32	3.74	2.579
Simpson's	0.96	0.79	0.170
Chao1	651.0	246.7	404.282

Table 2.8.
Overall Prediction accuracy of 10 replicates for the dataset comprised of ICoMM Arctic and Antarctic samples combined with other samples from the Arctic and Antarctic zones

		Actual	
		Arctic	Antarctic
Predicted	Arctic	79.7%	3.5%
	Antarctic	20.3%	96.5%
	Overall	100	100

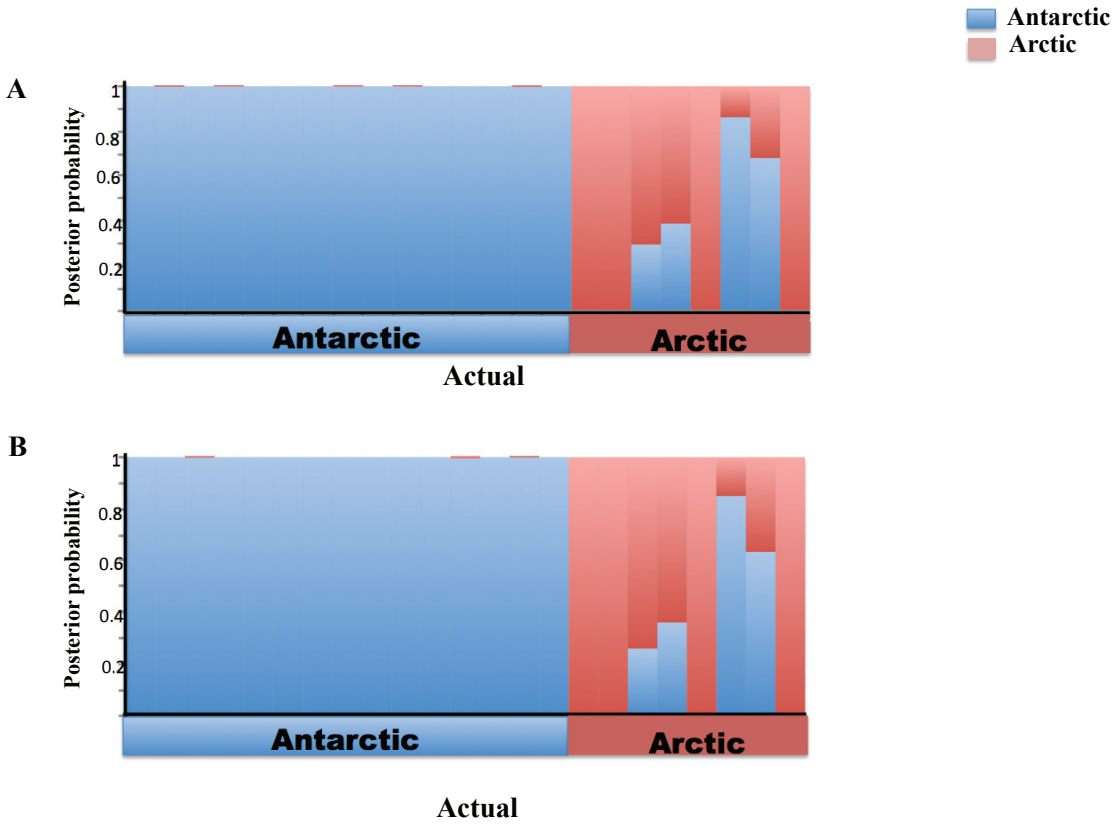

Figure 2.4: Bayesian prediction for Arctic and Antarctic samples combined with another sources of Arctic and Antarctic samples. (A) Prediction results from for the Arctic and Antarctic zones. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. Samples are plotted along the x-axis; for clarity the Arctic and Antarctic samples are grouped. The posterior predictions from just 1 of 10 cross-validation replicates (other were similar). (B) Prediction results derived from a longer run of the MCMC. As with panel (A), The posterior predictions are from just 1 of 10 cross-validation replicates. Note that there are fewer test samples than shown in Figure 2.3 because this Figure is based on a single replicate of the cross validation procedure, which was comprised of 1/3 of the total dataset.

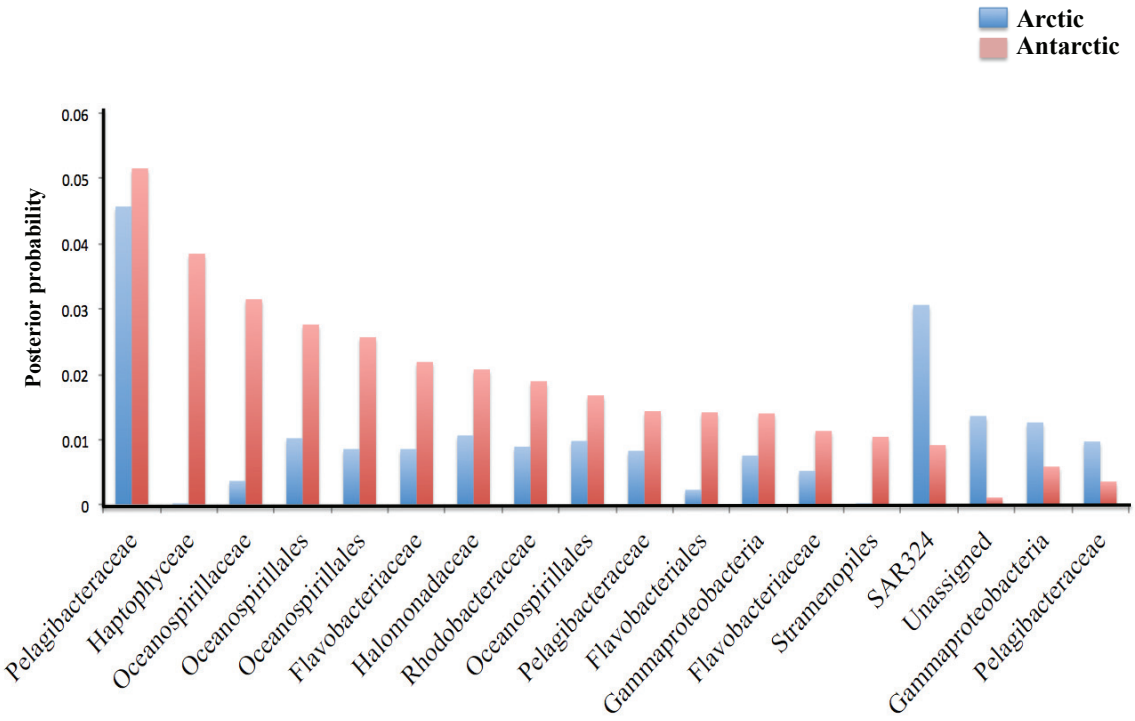

Figure 2.5 Comparison of predominant OTUs of the Arctic and Antarctic zone assemblages. Predominant OTUs were selected for both assemblages by using the >.01 posterior probability criterion. Note that because only limited taxonomic attribution was possible for these OTUs some taxonomic labels (e.g., Pelagibacteraceae) are repeated in this figure; each one of them represents a different OTU in this figure.

2.3.3 North Atlantic investigation: seasonal transitions among communities are not easily characterized for the entire ocean basin.

The temperate samples of the global analysis (section 2.5.1) entailed considerable heterogeneity (Appendices A1- A6) in terms of both environmental conditions and times of the year. Furthermore, previous studies of temperate sites (although not global in scope) suggest that within the water column at temperate sites there are strong and repeatable community transitions over a seasonal cycle (e.g., El Swais et al. 2015). I hypothesized that modeling of temperate sites might be improved by assembling a more focused dataset where all samples were from (i) the upper water column, and (ii) the North Atlantic temperate zone. My North Atlantic dataset (see methods for details) fits these criteria, and is comprised of 124 samples. These data were analyzed using a 2/3 training and 1/3 testing cross-validation design, with 5 random replicates. The model was trained according to season (winter, spring, summer and fall; see section 2.3.1 for more details how the samples were categorized), and was run using 10,000 iterations with a 5000 iteration burn-in interval. Surprisingly, the predictions of the trained model were very unsatisfactory, with and overall accuracy of just 39 %. Visual inspection of Figure 2.6 shows that none of the seasons were reliably predicted by the trained model. Indeed accuracy was just 33% for the fall, 38% for the spring, 25% for summer, and 60% for winter samples. Based on these results, I hypothesized that pattern of seasonal community transition might differ among the North Atlantic sites. This hypothesis is testable by narrowing the focus of the analyses even further, and attempting to characterize the temporal dynamics within a localized region within the North Atlantic. Since this had been investigated previously at a coastal inlet of the North-West Atlantic (Bedford Basin:

El-Swais et al. 2015). I choose to test this on the Eastern side of the Atlantic Ocean; the L4 station of the English Channel.

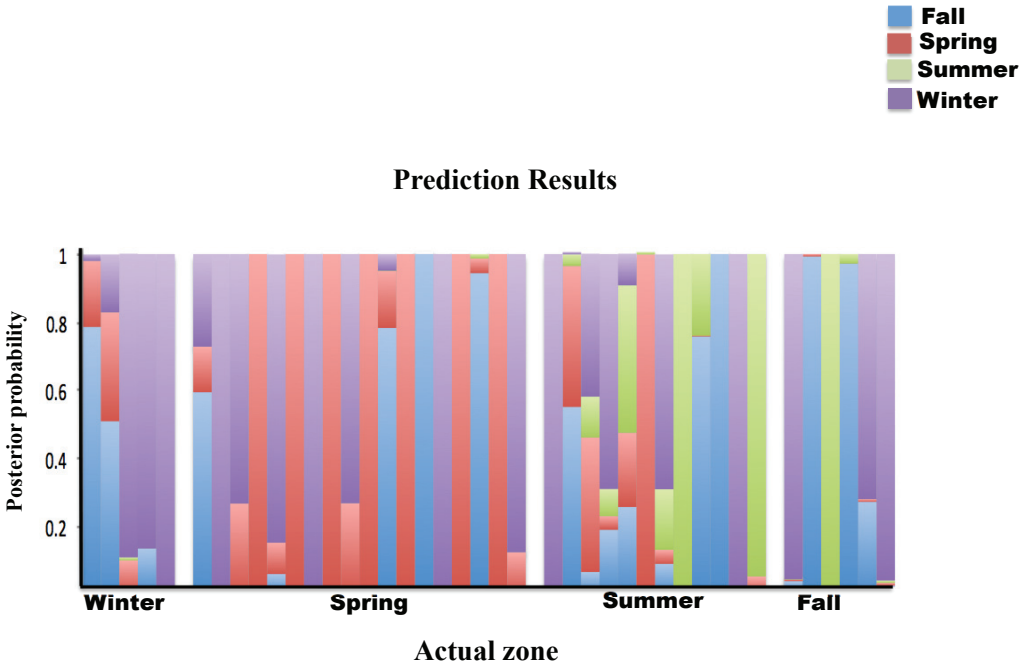

Figure 2.6: Bayesian prediction of temperate north Atlantic samples community structures. Prediction results for the four possible sample categories (winter, spring, summer and fall). The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. Samples are plotted in temporal sequence along the x-axis; for clarity the season is indicated on the x-axis instead of the sample ID. The posterior predictions are derived from just 1 of 5 cross-validation replicates (other were similar).

2.3.4 Seasonal variation at the L4 station of the English Channel: distinct communities can be resolved, and seasonal transitions are predicable

The L4 dataset is the only other single North Atlantic temperate site that has been extensively samples for seasonal variation. To test the hypothesis that distinct community transitions also occur at this site, I analyzed just the L4 data from temperate north Atlantic samples. This the L4 subset described in the methods (68 L4 samples in total). A model was trained to characterize samples taken only at the seasonal equinox (spring and fall) /solstice (summer and winter), and all remaining samples were treated as the test dataset. Analyses were based on a 1500 burn-in, and runs of 10,000 iterations. Figure 2.7 reveals that the seasonal community structures are recovered for the non-equinox samples (*e.g.*, the posterior distribution for samples taken in the spring is dominated by the spring equinox community structure). Moreover, the transitions between the seasonal structures might be occurring at points in time where biotic and abiotic factors measured at the L4 station are also undergoing major transitions. Each sample of L4 data was accompanied by a set of metadata consisting of some abiotic factors like (temperature, salinity, phosphate, Chlorophyll, etc.). Preliminary investigation of temperature, phosphate and chlorophyll changes over time did not reveal an obvious association with changes in community structure over time (data not shown). However, there are additional factors that can be added, and more comprehensive analyses are warranted.

2.3.5 Seasonal variation at the L4 station as a model for the North Atlantic: L4-derived community structures do not generalize to the North Atlantic

The model was re-trained on L4 equinox/solstice samples and applied to the North Atlantic dataset used in section 2.3.3 above. Results (Figure 2.8) were not as good as those obtained when the model was applied to just the L4 test data (Figure 2.7). Overall accuracy of this model for the North Atlantic was 55%. This was an improvement over predictive accuracy for the temperate-seasons in the global dataset (37.7%), despite the fact that the training sample was reduced. However, performance of this L4- trained model was substantially lower than the predictive accuracy obtained when it was applied only to the L4 test data (84.6%). Taken together, these results suggest that season community transitions can be predicted for North Atlantic water column sites (L4 in the Eastern Atlantic, and Bedford Basin in the Western Atlantic), but that the temporal dynamics of such transitions differ among sites and are not well modeled when different sites are pooled into a single dataset. This is because seasonal variation occurs at all temperate sites, but the transitions occur at different times at different locations (Caporaso et al,2012). It seems that the entire North Atlantic has too much temporal heterogeneity for this modeling approach.

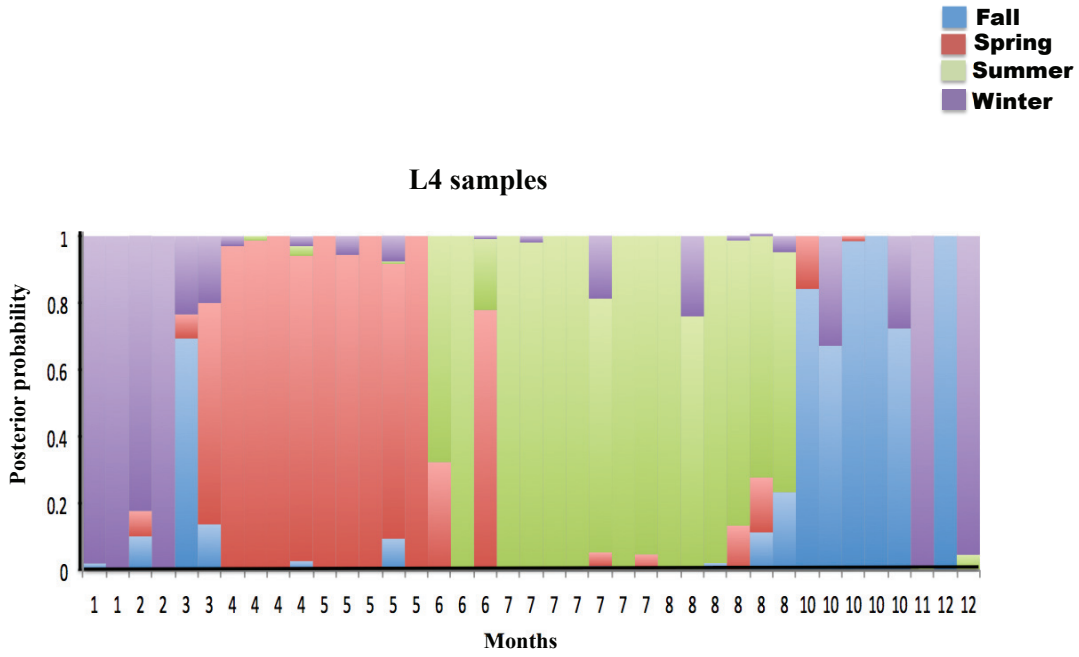

Figure 2.7: Predictions results for L4 samples using a model trained on the equinox/solstice samples from that site. Prediction results for L4 samples from the north Atlantic samples. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. Samples are plotted in temporal sequence along the x-axis; for clarity the months (as a number) is indicated on the x-axis instead of the sample ID.

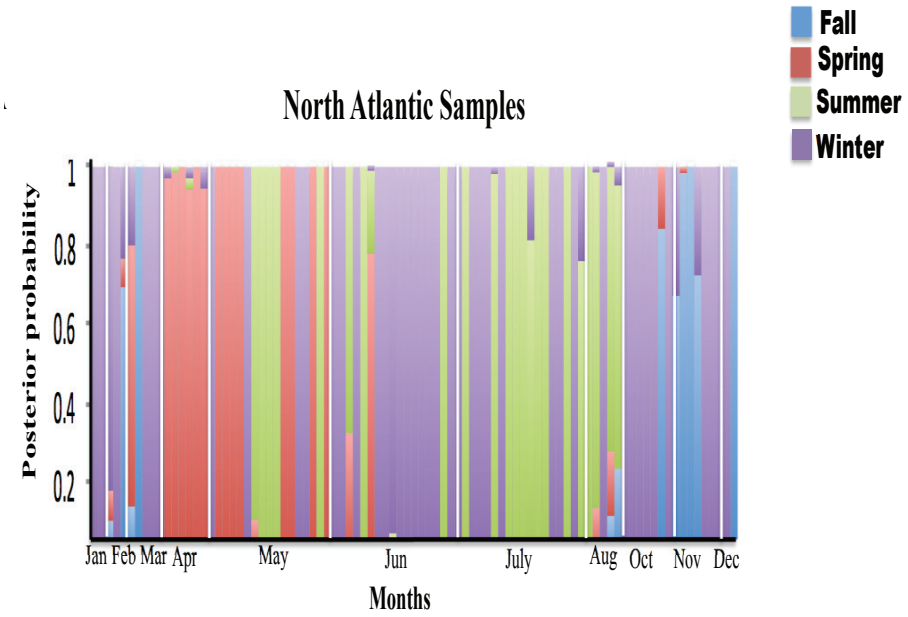

Figure 2.8: The predictions results of north Atlantic samples using equinox/solstice L4 samples to train on and test the rest of north Atlantic samples. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. Samples are plotted in temporal sequence along the x-axis; for clarity the months is indicated on the x-axis instead of the sample ID

2.3.6 Seasonal variation at the L4 station as a model for the English Channel: L4-derived community structures do provide some predictive value for the English Channel

Following the same approach used in section 2.3.5, a model was re-trained on L4 equinox/solstice samples and applied to the remaining samples in the expanded English Channel dataset (99 samples in total). As the model was trained on only the L4 equinox/solstice samples for each season, those samples were excluded from the test data. Thus the test data was comprised of 72 of samples from L4 and nearby in the English Channel. Analyses were based on a 1500 burn-in and 10,000 iterations. Although not as good as for the L4-only test data, the predictions were generally good with overall accuracy of 55% (Figure 2.9), with seasonal shifts between communities clearly visible via the posterior distribution. Where the model inferred mixtures of assemblages, they tended to be mixtures of adjacent seasonal assemblages; i.e., winter with spring, spring with summer, and summer with fall (Figure 2.9). Of course, I trained on a subset of the L4 samples, and the test data contained many other samples from L4, so it is reasonable that good predictions were obtained.

Taken together, the results suggest that seasonal transitions among communities within the temperate North Atlantic water column are repeatable; this finding supports the view of Becking (1934) that environmental factors drive biogeographic patterns (in this cases season-specific biogeographic patterns). Furthermore, assuming that the seed bank hypothesis is correct (Gibbons et al. 2013) my results are consistent with the notion that the assembly of these seasonal communities is based on growth of bacteria that are present at this location throughout the year, even if at times their abundances are very low. The finding that the L4-based predictive model was not as good for nearby (North Atlantic) sites indicates that different temporal dynamics may be happening at those other

sites. Presumably, all temperate sites go through similar transitions but at different times. However, the specific assemblages may also differ. Finally, the finding of predicable transitions between assemblage structures by El-Swais et al (2015) at an Western North Atlantic site, taken together with our results for an Eastern North Atlantic (L4) site, suggests that environmental driving of seasonal community assembly is common to the temperate North Atlantic, and may be a general explanation for bacteria biogeographic patterns in the North Atlantic. It remains to be seen if there are broad similarities between the environmental drivers that may be operating in the Eastern and Western Atlantic.

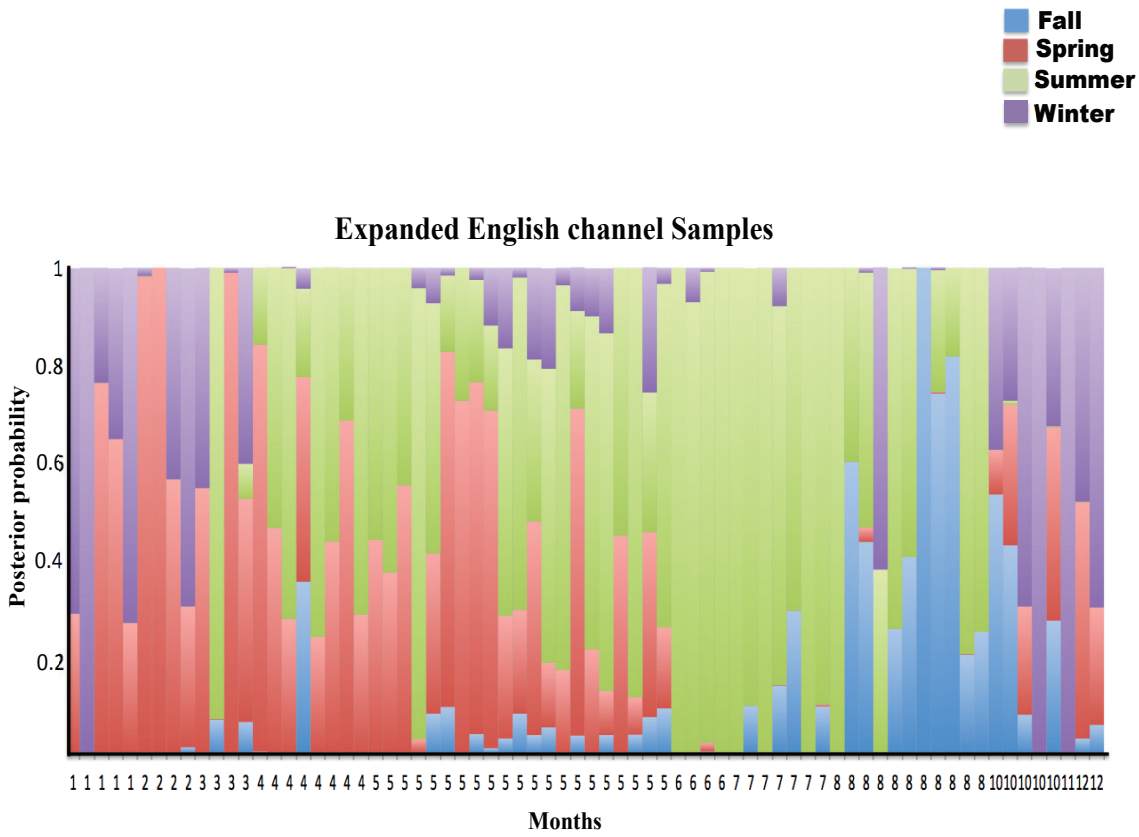

Figure 2.9: The prediction results for the expanded English Channel samples. The height gives the posterior probability for assemblages. The assemblage with the highest posterior probability is the best prediction for that sample. Samples are plotted in temporal sequence along the x-axis; for clarity the months (as a number) is indicated on the x-axis instead of the sample ID.

CHAPTER 3:EVALUATION OF ALTERNATIVE ANALYTICAL STRATEGIES FOR BIOMICO

3.1 INTRODUCTION

Biological inference under BioMiCo requires statistical inference of posterior distributions via Markov Chain Monte Carlo (MCMC). MCMC is a powerful, but complex, means of sampling from a probability distribution. In the case of BioMiCo, it is used to sample from the posterior probability distribution of both the OTU mixture weights and the assemblage mixture weights (see Chapter 1 for an overview of the structure of the BioMiCo model). The problem is that running an MCMC to get a good result, either for BioMiCo or for other complex modeling frameworks, can sometimes be a challenge. There are several factors involved; chief among these are (i) length of the MCMC run, (ii) the amount of "burn-in" to discard, and (iii) the error associated with estimation of the posterior probability distribution obtained via MCMC and (iv) the variability of results among separate runs due to such errors. The objective of this chapter is to evaluate how those factors could influence results obtained from BioMiCo.

3.1.1 Brief overview of MCMC and the challenges it poses

Markov chain Monte Carlo is a random sampling algorithm invented in the 1940's by physicists at the Los Alamos National Laboratories, one of the few places that had computers at that time. The method was used to solve problems that were impossible to solve using the conventional approach based on deterministic mathematical models. Bayesian statisticians faced an analogous situation; some Bayesian inferences could only be done by using MCMC. MCMC generates random draws from a target distribution (for

Bayesian inference, this is the posterior distribution) and provides a way to construct a "nice" Markov chain such that the stationary distribution for the Markov process is equivalent to the target distribution. The goal of MCMC is to draw samples from a probability distribution without knowledge of its exact height at any point, and it does so by "wandering around" on that distribution (i.e., sampling that distribution) such that the amount of spent time in every location is proportional to the distribution height. This is possible because a Markov Chain is a "memoryless" random process; the probability distribution of a future state depends only on the current state of the process and not the past.

Just because a Markov chain can be constructed with the desired equilibrium properties, does not mean it will be easy to run that chain until it has converged to its stationary distribution. The samples from the Markov chain are referred to as "steps", and it can be difficult to determine when enough steps have been taken to ensure that it has converged. There exists many algorithms to perform MCMC, and they have been applied to a wide variety of problems. For some problems, the stationary distribution can be reached quickly, whereas for others it can take a very long time. Furthermore, MCMC can only ever approximate the target distribution, so the amount of error in the approximation that can be tolerated is also a factor. The starting position of the MCMC plays a role in the amount of error associated with the approximation.

In principle, convergence can be achieved for any parameter if the chain has been run long enough. The simplest way is an extremely long run, however, long runs are sometimes difficult to achieve because computers may fail after several hours, days or months of operation. One suggestion for collecting enough samples for a "long enough run" is to run multiple short chains and combine the results. The lack of a good answer

with a single long run is an indication that it is not possible to obtain a good answer with the combination of many short runs having the same total number of samples.

To minimize the impact of the starting point on the error, users of MCMC will often discard what they call the "burn in". The burn in is necessary because BioMiCo starts with in an initial state of equally likely probabilities for each parameter (i.e., the symmetric Dirichlet prior probabilities). The probabilities for the factor variables (the π, of Shafiei, et al. 2015, see also Chapter 1), the probabilities defining the contributions of each assemblage to each factor (the θ in that paper), and the probabilities of each OTU to each assemblage (the ϕ of that paper) are initially equal. As such, the initial state of system is likely far from the final state. It will take some time for the system to converge on something that approximates the desired posterior distribution. This burn-in period does not represent the desired "nice" Markov chain with a stationary distribution that is equivalent to the target distribution. Unfortunately, it is not possible to know (to predict) the necessary burn-in time because it depends on the difference between the prior and posterior distributions, as well as the number of parameters in the model.

3.1.2 General Issues with BioMiCo

Inference about community structure relies on the capacity to reliably estimate the posterior mixture of assemblages within a community, and the mixture of OTUs within an assemblage. BioMiCo uses MCMC to estimate those posterior probabilities. There are four problems that arise from this. First, as the contribution of OTUs to assemblages is probabilistic, there are no hard boundaries between the different assemblages of OTUs. Thus, assemblages must be compared according to the notion of "predominant OTUs" and the definition of such OTUs is subjective. Second, the posterior probabilities that are

used to decide if an OTU is predominant within a given assemblage will have errors that arise from the finite number of samples from the MCMC chain, and the starting point of the chain. Third, the design of the testing and training phases of the analysis could impact the error. Fourth, the number of assemblages in the model must be fixed beforehand. Since the optimal number will not be known for real data, the practice is to set the number to be larger than what the user thinks might be optimal (i.e., intentionally set too high), and rely on the sparse Dirichlet priors to minimize variance and enhance interpretability of the results. Setting this too high might also contribute error to the inferred posterior probabilities. In this chapter I investigate each of these issues, in turn, within separate sections (3.2 to 3.5).

3.2. Alternative Methods To Identify Predominant OTUs

Here, I investigate four alternative criteria for defining the "predominant" OTUs within a given assemblage of microbes. All four are based on the posterior probabilities inferred by using BioMiCo. The first is to use a posterior probability value ≥ 0.01 as the criterion (hereafter, the 0.01 criterion). Note that this was the criterion that was used in Chapter 2, and has been used in the literature (El-Swais et al. 2015). The three alternative criteria include: (*i*) the OTUs with the highest probabilities that sum to the 95% posterior density (hereafter, the 95% criterion), (*ii*) the OTUs with the highest probabilities that sum to the 50% posterior density (hereafter, the 50% criterion), and (*iii*) all OTUs above the inflection point of the posterior distribution. Note that using of the inflection point requires subjective assessment of the shape of the posterior distribution, but it has been previously suggested as a possible criterion (Shafiei et al. 2015).

I compared the impact of these different criteria on the choice of predominant OTUs in the Polar dataset (Arctic and Antarctic), and in the L4 dataset (see Chapter 2 for additional details). Recall that each of these datasets was originally run with L=25 assemblages, which is more than is necessary to describe these data (Chapter 2), as I found that there was just one assemblages with high posterior probability for the Arctic and one for the Antarctic zones. Hence, I investigated the predominant OTUs in just those two assemblages. As I had run 10 replicates of 2/3 training with 1/3 testing (see Chapter 2), results derived from 2/3 training were compared across the 10 replicates. For the L4 dataset, OTU composition was investigated for the dominant assemblage in each of the four seasons (fall, spring, summer, winter). For this assessment there is no replication, as training was carried out on just the equinox/solstice samples (rather than random subsets).

Table 3.1 presents the results for the 0.01 criterion, the 95% criterion and the 50% criterion for the Arctic and Antarctic samples. In all cases, the 0.01 cut-off was the most conservative. Under this criterion, the 10 replicates for the Antarctic assemblage were comprised of 13 to 16 OTUs (mean: 14.8, SD: 0.92), and the 10 replicates for the Arctic assemblage were comprised of 5 to 13 OTUs (mean: 8.50, SD: 2.84). This was far less than the 50% criterion for the Antarctic sample with 39 to 47 OTUs (mean: 43.50, SD: 3.21) and for the Arctic sample with 75 to 111 OTUs (mean: 92.90, SD: 10.86). Both of these criteria were less than the 95% posterior probability criteria criterion for the Antarctic sample with 746 to 843 OTUs (mean: 796.20, SD: 40.02) and for the Arctic sample with 976 to 1202 OTUs (mean: 1120.60, SD: 72.49). Note that the 0.01 cut-off represented between 30 to 35% of the posterior density for the Antarctic replicates (mean: 33.10%, SD: 1.37) and between 11 to 21% of the posterior density for the Arctic

replicates (mean: 15.10%, SD: 3.47). Given the way the criteria work, the OTUs that are within the 0.01 criterion are a subset of those in the 95% and 50% criteria. In addition, the number of OTUs for the Arctic sample consistently exceeds those of the Antarctic sample (using all criteria), and these samples also have higher species richness and diversity (Table 2.7).

The 0.01 criterion is the most conservative. Generally, the fact that the 0.01 criterion produces a posterior density less than the 50% (or 95%) implies that these samples consist of a small number of "higher probability" OTUs. That is, for the Antarctic samples, the average proportion for those exceeding 0.01 was 0.022, and for the Artic samples, the corresponding value was 0.018. Note that the notion of "higher probability" OTUs here is relative to the expectation derived from a uniform distribution for OTU posterior mixture weights. For comparison, if we assume a uniform posterior distribution (totally uninformative) the probability of each OTU is just 0.000075 (which is 1/13312) for the Arctic and Antarctic samples, respectively. Thus, the inferred mixture weights for OTUs above the 0.01 cutoff are relatively large, and indicative of informative structure within the data. From another perspective, the 95% criterion required nearly 1000 OTUs implying the inclusion of many lower probability OTUs.

The same procedure was applied to the L4 data, as shown in Table 3.2. For the summer, fall and winter only 17 OTUs exceeded the 0.01 criterion, accounting for between 48% and 54% of the posterior distribution. The spring was somewhat higher with 22 OTUs and 62% of the posterior distribution. Across all four seasons, the mean number of OTUs was 18.25 (SD: 2.50), and the mean density was 53.25 (SD: 6.40). The 50% criterion ranged from 12 to 18 OTUs (mean: 15.00, SD: 3.46), while the 95% criterion ranged from 192 to 283 (mean: 228.25, SD: 44.10). Note that in contrast to the

previous Antarctic and Arctic analyses, the 0.01 criterion is essentially the same as the 50% criterion. Note that the 50% and 95% posterior probability criterion identify far fewer OTUs for the L4 samples than for the previous Arctic and Antarctic samples. This implies that this region is dominated by OTUs that have higher probabilities.

Table 3.1 The effect of different definitions of "predominant OTU" on the size of the primary assemblage (number of OTUs) over 10 training replicates for the Antarctic and Arctic samples**.**

		No. OTUs in the primary assemblage	
Training replicate	**Criterion**	**Antarctic**	**Arctic**
1	95%	832	1202
	50%	45	100
	0.01	14 (34%)	6 (12%)
2	95%	840	1201
	50%	47	98
	0.01	16 (32%)	7 (13%)
3	95%	747	1055
	50%	40	84
	0.01	15 (34%)	10 (19%)
4	95%	831	1062
	50%	45	81
	0.01	15 (33%)	11 (18%)
5	95%	746	976
	50%	39	75
	0.01	15 (34%)	13 (21%)
6	95%	843	1120
	50%	47	92
	0.01	13 (30%)	10 (16%)
7	95%	810	1129
	50%	45	90
	0.01	16 (33%)	11 (17%)
8	95%	750	1146
	50%	46	103
	0.01	15 (35%)	5 (11%)
9	95%	774	1190
	50%	39	111
	0.01	14 (33%)	5 (11%)
10	95%	789	1125
	50%	42	95
	0.01	15 (33%)	7 (13%)

Note: values in parenthesis are the posterior density for the OTUs identified using the 0.01 criterion

Table 3.2 The effect of different definitions of "predominant OTU" on the size of the primary assemblage (number of OTUs) for the L4 dataset (summer, fall, winter and spring) samples**.**

	No. OTUs in the primary assemblage			
Criterion	**Winter**	**Spring**	**Summer**	**Fall**
95%	283	192	193	245
50%	18	12	12	18
0.01	17 (48%)	22 (62%)	17 (54%)	17 (49%)

(Note: values in parenthesis are the posterior density for the OTUs identified using the 0.01 criterion)

Among replicates variability increased with the inclusion of more predominant OTUs, presumably because the additional OTUs have posteriors that are more sensitive to estimation errors. The impact of these errors can be seen in both the roughness of the posterior distributions (Figure 3.1, and described above), and the variability of that roughness among replicates (data not shown). Based on these results, I suggest that a conservative criterion should be employed for identifying predominant OTUs, which will help to mitigate the effect of the estimation errors.

The inflection point method has been used previously (Shafiei et al. 2015). However, this method was impractical for the Arctic and Antarctic data. The problem was that the shape of posterior distributions was “rough” in the region where the inflection point should have been, meaning that (at a local level of analysis) there were multiple points where the shape of the curve changed (sometimes dramatically). More generally, there was no inflection point (i.e., a point where the second derivative went to zero). The curve resembled a negative exponential or an inverse function. This was the case for all replicates, but only a single replicate for the Arctic and Antarctic is shown in Figure 3.1. This result suggests that there is considerable uncertainty in the estimates of the posterior probabilities in this region. For the L4 data, the curves are much smoother, but there is still no clear inflection point (Figures 3.2). Normally, an inflection point would delineate a group of microbial species "equally" high probabilities from a group of species with "equally" low probabilities, and such a point would serves as a good break. The lack of a good inflection point likely implies that estimation errors are too large in this part of the curve to make this a consistent method. Hence, the inflection point method is not considered further.

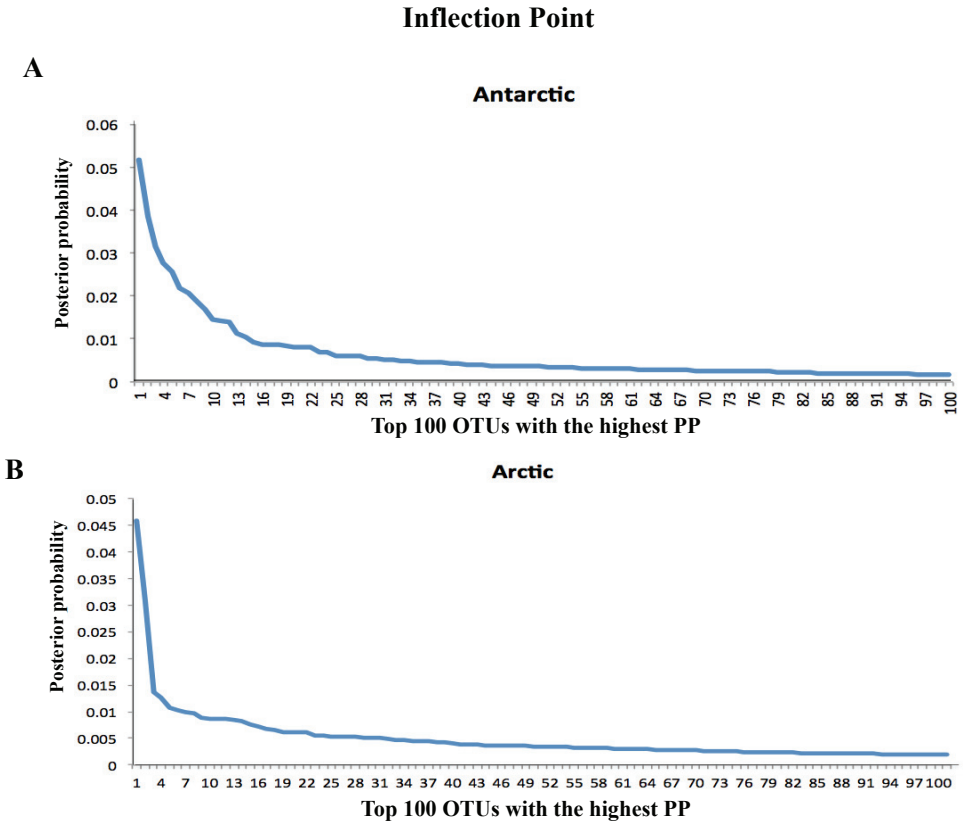

Figure **3.1** The posterior probability (y-axis) as a function of the top 100 OTU with the highest posterior probabilities (x-axis) for one of the Arctic and Antarctic replicates (others were similar). Note the "roughness" of the curve near where the inflection point might be located.

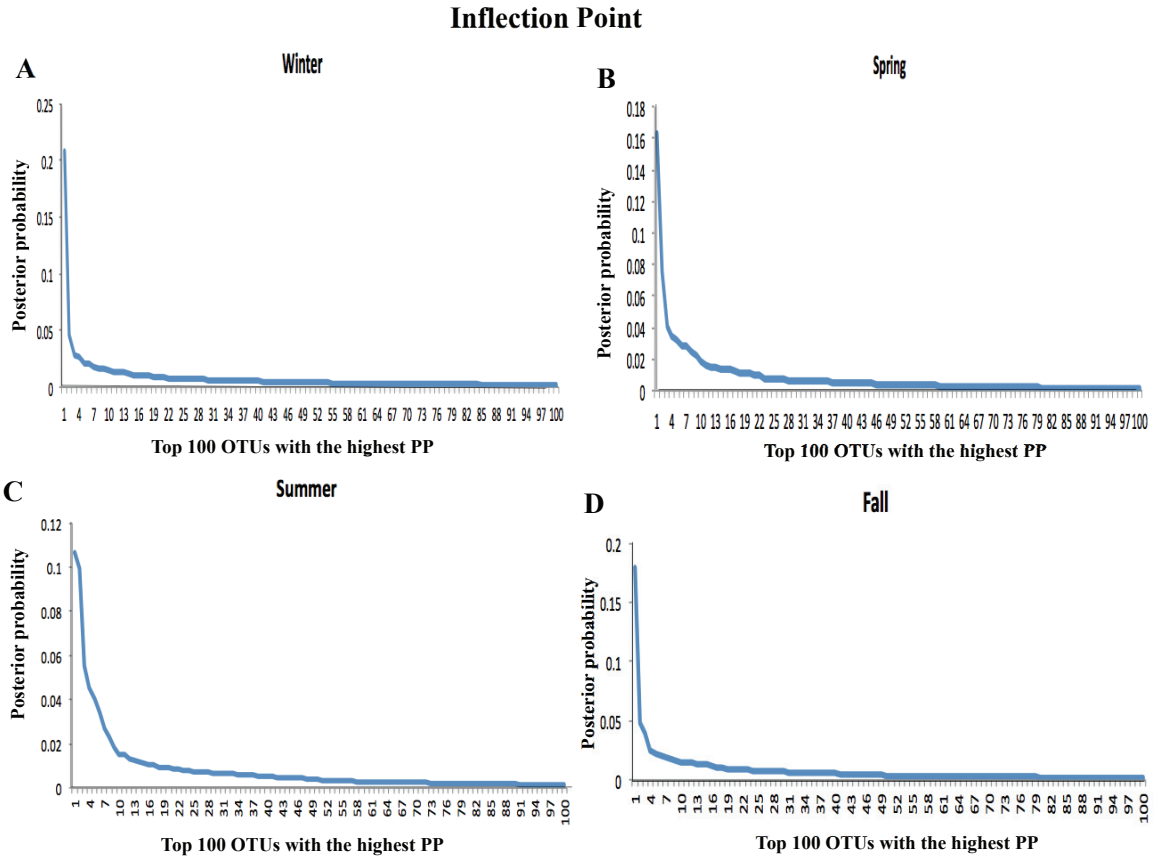

Figure 3.2 The posterior probability as a function of the top OUT with the highest posterior probabilities for the winter (A), spring (B), summer (C), fall (D) samples.

3.3 The Impact of Burn-in And Number of MCMC Iterations

Inference of posterior distributions under BioMiCo involves integrating out the latent variables θ, ϕ, π (described in Chapter 1) and sampling from the posterior distribution for OTU assignments to assemblages, and assemblage assignments to the factor variables. This is carried out using Gibbs sampling. In short, given a full set of variables, each state of the Markov chain is sampled for an assemblage and factor assignment for one OTU (indexed by *i*) conditioned on the current state for all other variables in the model. New values for the assemblage and factor assignments are then drawn from the conditional distribution for this OTU, and those values are updated within the model. The updated values are then fixed, and the process is repeated for a different (randomly selected) OTU. One "*iteration*" of Gibbs sampling is complete when the process has been carried out for each OTU in each microbiome sample.

Because the start-point of the MCMC involves random initialization of the model variables, posterior probabilities derived from the beginning iterations of the MCMC do not represent samples from the stationary distribution. This is why the starting point of the MCMC contributes error to the estimate of the target distribution, and is typically discarded. The discarded iterations are referred to as "*burn-in*", and the user determines the number of burn-in iterations. Following the burn-in, the MCMC is run for a fixed number of iterations. Due to the conditioning, successive iterations are not independent samples from the target. For this reason, the target distribution is estimated by sampling the MCMC at *intervals*. In this study, the sampling interval was 500 iterations. In real data analyses, the MCMC is typically run multiple times, and the posterior distributions are checked for concordance. The separate runs of the MCMC are referred to as "*replicates*".

To investigate the impact of the burn-in, all the data were put into the training set (there were no test sets). The impact of the size of the burn-in for the posterior distribution was examined for various burn-in lengths starting with no burn-in. The effect of the burn-in length was investigated by computing the posterior distribution of the assemblages after the first interval (500 iterations after the burn-in) and also at additional intervals until well-defined assemblages were observed. This procedure was applied to both the polar data (70 samples and 2 factor labels) and the L4 equinox/solstice data (68 samples and 4 factor labels).

Examining the intervals using a zero burn-in allows for an investigation of changes in posterior distributions with increasing intervals and should aide in determining an appropriate burn-in period. For the Arctic and Antarctic dataset the first interval (500 iterations) for the zero burn-in case shows the assemblage distribution was uniform; that is, there were no dominant peaks (Figure 3.3A). Given that the starting assemblage distribution is uniformly distributed implies that 500 iterations is not sufficient to even approach the ideal solution, and that burn-in should exceed 500 iterations. Examination of the second interval (after 500 additional iterations) shows two well defined assemblages for the Arctic and Antarctic (Figure 3.3B). This implies that the assemblage posterior distribution has begun to stabilize by 1000 iterations. To attempt to fine-tune the point that this occurred, burn-in length was increased, starting with a burn-in of 100. Using a burn-in of 100 iterations, the first interval (still 500 iterations) shows two well-defined assemblages, which suggests that the assemblage distribution has begun to stabilize at 600 iterations (100 iterations of burn-in plus 500 iterations in first interval). This suggests a burn-in of 600 iterations is a minimum number for this dataset (however, this is unlikely to generalize to other datasets).

A similar investigation was applied to the L4 data. At zero burn-in at interval 1 (500 iterations) the assemblage distribution was uniform; that is, there were no dominant peaks and at interval 2 it was no longer a uniform but still there were no distinct peaks. This implies that a burn-in of at least 1000 might be necessary for these data. At interval 3 (after 500 additional iterations) I obtained well-defined assemblages for each season for L4 data. This implies that the assemblage posterior distribution has begun to stabilize by 1500 iterations. I found that discarding a burn-in of 1000, and running the chain for just 500 iterations also produced well-defined assemblages (Figure 3.4).

It is important to note that the observation of well-defined assemblages does not necessarily mean that a stable estimate of the posteriors assemblage distribution has been obtained (it may take many more iterations to achieve the desired level of error in the approximation to the target distribution). Longer runs, with replication, are required to determine this. We recommend that real data analyses should be run longer than just 500 iterations post burn-in. What we have shown is that (i) iterations associated with the burn-in interval can have a large contribution to the error in the posterior distribution, and (ii) the size of the burn-in to discard is somewhat dependent on the data and the model.

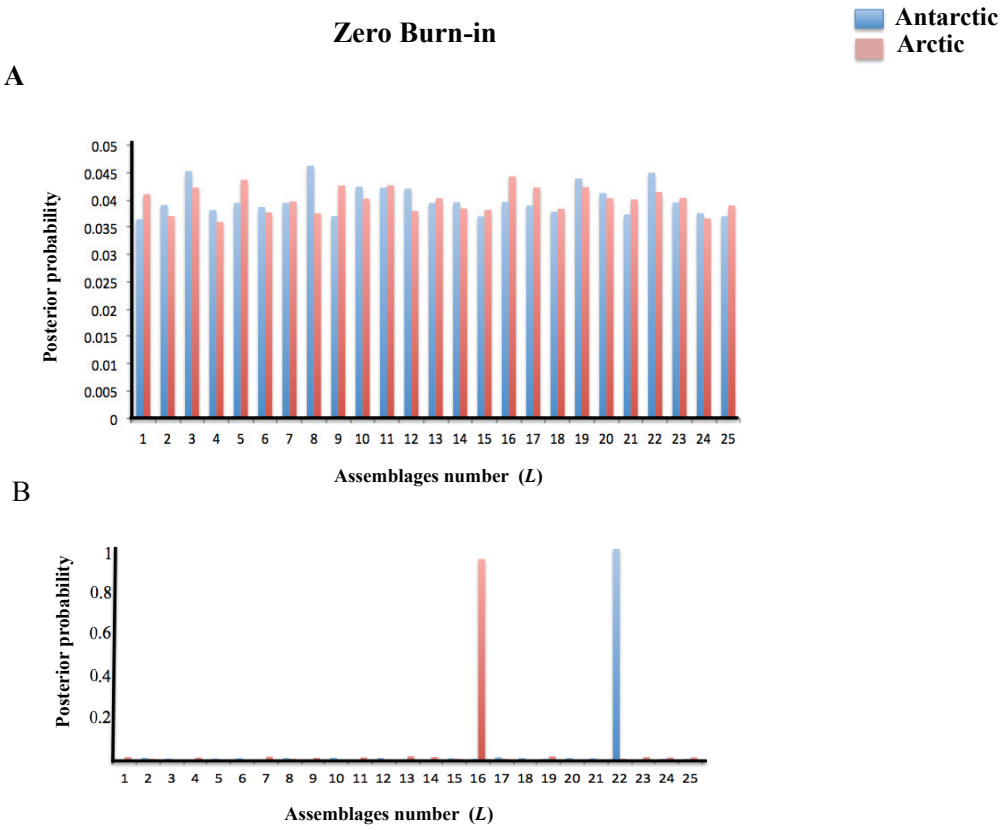

Figure 3.3 (A) The assemblage distribution from zero burn-in at interval 1. (B) The assemblage distribution from interval 2 with zero burn-in. The total is 1000 iterations.

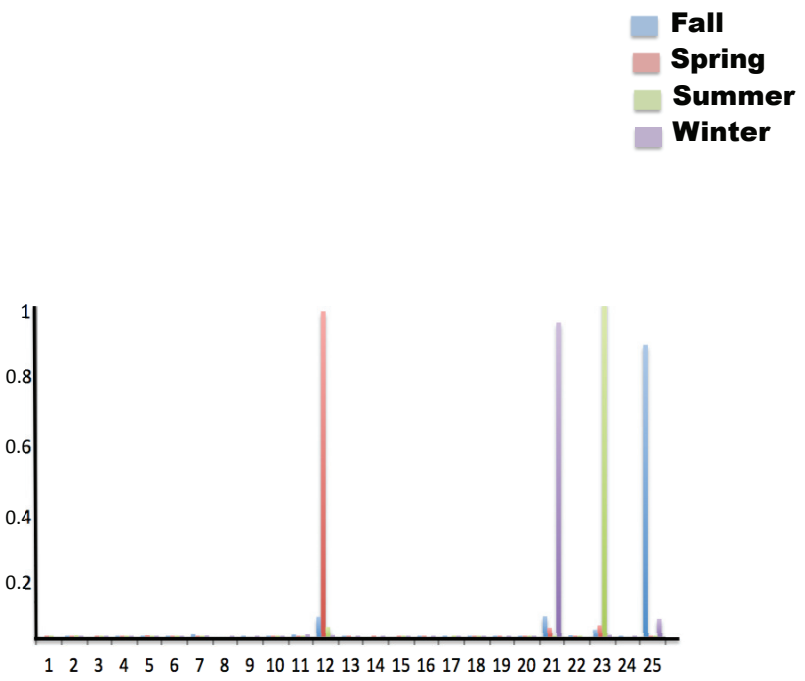

Figure 3.4 The distribution of seasonal assemblage for L4 data After a burn-in of 1000, the MCMC was run for 500 iteration.

3.4 The Impact Of The Design Of The Training And Testing Phase

The goal of this analysis is to examine the different designs of training and testing on prediction level. In chapter 2 I trained on 2/3 of the Arctic and Antarctic samples and tested the remaining 1/3 of the samples, which I will refer to as 2/3:1/3. The samples were assigned to training and testing sets randomly over 10 replicates. Two additional methods of assigning data to training and testing sets were also investigated: (i) training on 90% of the samples and testing the remaining 10%, which I will refer to as 90:10, and (ii) training on 50% of the samples and testing the remaining 50%, which I will refer to as 50:50. Like in Chapter 2 the samples were assigned to training and testing sets randomly and replicated 10 times. To rule out the impact of burn-in and iterations, a long burn-in (1500 iterations) and large number of iterations (10,000) was used.

In general, the predictions of BioMiCo when 2/3:1/3 strategy was used were good (Figure 3.5A). Similar patterns were seen between the replicates. Using 90:10 also showed good predictions for Arctic and Antarctic in all the replicates, and was quite similar to the 2/3:1/3 (see Figure 3.5B) with an overall accuracy of 100% for Arctic and 96% for Antarctic samples for all the 10 replicates. However, when 50:50 was used the number of incorrect predictions was more than what was seen in the 2/3:1/3 and 90:10 strategies (Figure 3.5C). The overall prediction accuracy for 50:50 training and testing design for the 10 replicates is 96% for Antarctic and 78% for Arctic samples. It is noticeable that the more samples that are used to train, the better the prediction, but that using more than 2/3 would not likely generate further improvement.

I did not apply the three strategies for assessing the impact of training and testing design to the L4 datasets because for the L4 data training was conducted with the equinox/solstice samples, while testing used the rest of the data (see Chapter 2).

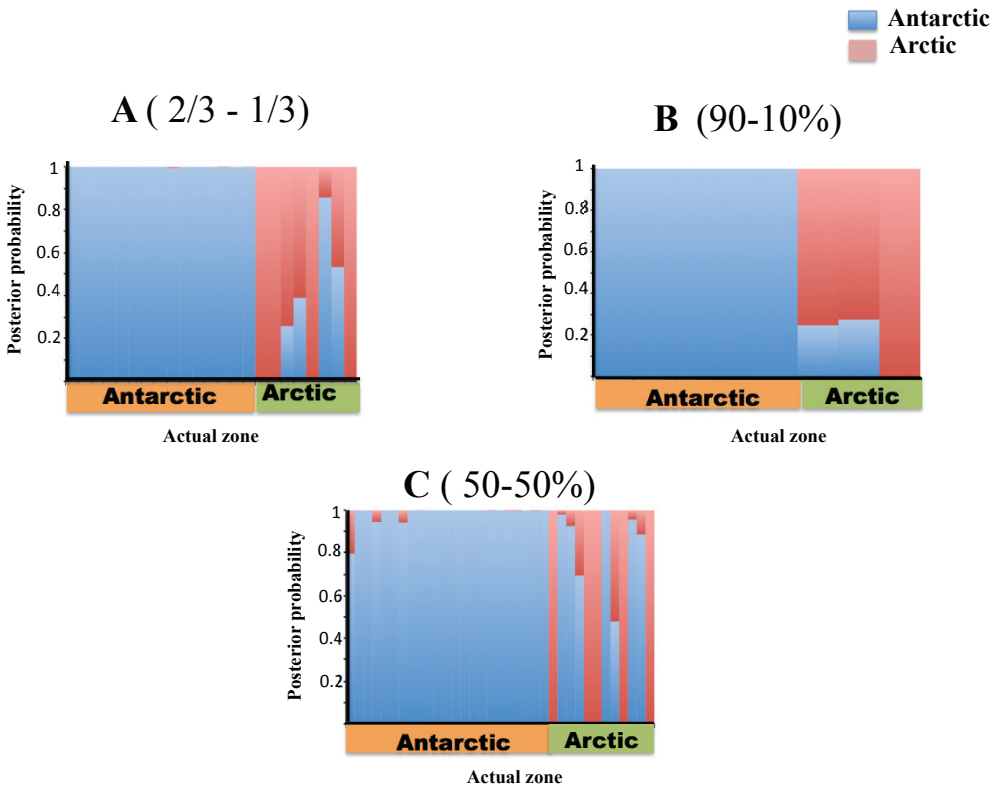

Figure 3.5 Prediction results of different design of training and testing for Arctic and Antarctic samples. (A) training 2/3 and testing 1/3, (B) training on 90% and testing 10%, and (C) training on 50% and testing 50%. Results are shown for one of the 10 replicates; similar patterns were seen in the other replicates.

3.5 The Impact of Number Of Assemblages On The Posterior Distribution

The purpose of this analysis was to understand how setting a pre-specified number of assemblages (L) might have impacted the final posterior distribution. In my previous real data analyses (Chapter 2) I set the number of assemblages to 25, assuming that value would be more than sufficient to capture the community structure. To examine the impact of assemblage number (L) the Arctic and Antarctic samples were rerun using different values of L to see if this made any changes in number of predicted major assemblages and size of the set of predominate OTUs. L values of 5, 10, 20 and 40, were evaluated. The number of predicted major assemblages for an L of 25 was compared with those found at L values of 5, 10, 20 and 40 for the Antarctic, Arctic, and L4 data. In addition, the size of the set of predominant OTUs under the previous 3 criteria (i.e., 95%, 50% and PP > 0.01 criteria) was also examined.

For the Arctic and Antarctic samples all L values yielded two predominant assemblages, one for each label (as in my previous analyses). The size of the set of predominant OTUs under the 0.01 criterion were unaffected by the value of L (Table 3.3). Under this criterion, all Antarctic samples had 14 predominant OTUs (posterior density between 29% and 31%), and the Arctic samples had 7 or 8 (posterior density between 12% and 14%).

In contrast, the size of the set of predominate OTUs was very sensitive to the value of L under the 95% criteria (Table 3.3). The size of predominant set for the Antarctic samples decreased from 1583 (L=5) to 800 (L=40). The same effect was observed for the Arctic samples; the size decreased from 2223 (L=5) to 1129 (L=40). The results for L=20 and L=40 were comparable to the average result for the real data analyses where L was set to 25 assemblages (Antarctic mean = 796.2 OTUs Arctic mean= 1120.60

OTUs). The sensitivity of the 50% posterior probability criterion was intermediate between the other two criteria (Table 3.3). The Antarctic mean size under the 50% criterion and *L*=25 was 43.50 OTUs and Arctic mean size was 92.90 OTUs; these are close to, but a little less than, the size when *L*=20 and *L*=40 in Table 3.3. These results also support using the more conservative criterion for identifying the predominant OTUs (0.01), as it appears to be less sensitive to errors in the estimation of the posterior distribution.

I expect that small number of assemblages will tend to soak up more random errors than when there are larger numbers of assemblages in the model. Nonetheless, the results in Table 3.1 imply that the dominant OTUs remain consistent regardless of number of assemblage in the model, and that a reasonable estimate of assemblage composition can be made as long as the value of *L* is not too small (Tables 3.1 and 3.3 taken together). Graphically, the assemblage distributions show little change from running with different number of communities in Arctic and Antarctic samples; the assemblage distribution when *L is* 5 (Figure 3.6A) is quite similar to that when *L* is 40 (Figure 3.6 B). The peaks for the different *L* values were stable in that both showed two predominate peaks, one for the Arctic and the other for the Antarctic.

Table 3.3 The effect of the number of assemblages (*L*) on the size (number of OTUs) in the predominate assemblage for the Arctic and Antarctic samples.

		No. OTUs in primary assemblage	
Assemblages (*L*)	**OTU criterion**	**Antarctic**	**Arctic**
40	95%	800	1129
	50%	45	97
	0.01	14 (31%)	8 (14%)
20	95%	880	1199
	50%	46	100
	0.01	14 (31%)	7 (13%)
10	95%	1104	1595
	50%	49	111
	0.01	14 (31%)	7 (12%)
5	95%	1583	2223
	50%	54	126
	0.01	14 (29%)	7 (12%)

Note: values in parenthesis are the posterior density for the OTUs identified using the 0.01 criterion

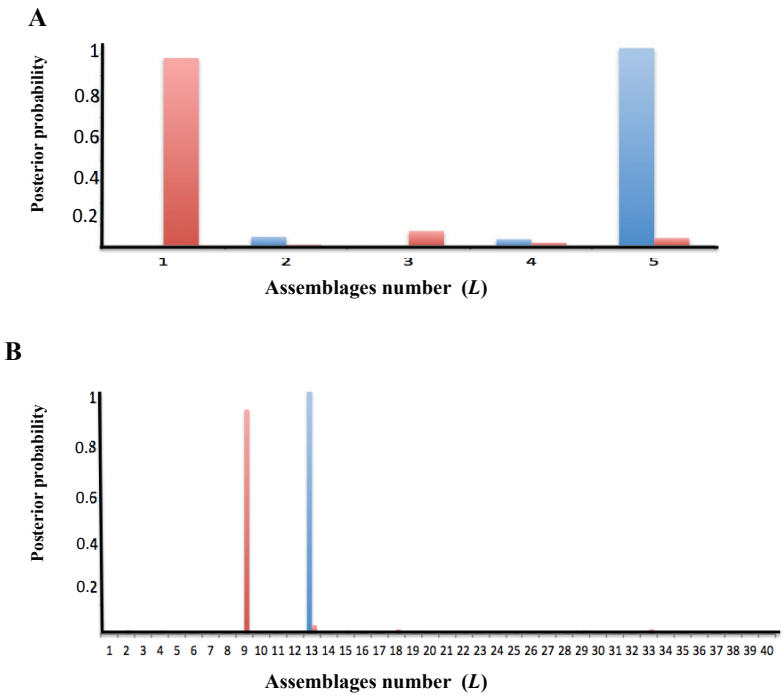

Figure 3.6 The assemblage distribution of Arctic and Antarctic zones from running different number of assemblages. (A) The assemblage distribution when $L = 5$. (B) The assemblage distribution when $L = 40$.

Similar results were obtained for the L4 dataset (Table 3.4). Again, the size of the assemblages, in terms of the number of predominant OTUs, decreased with increasing values for *L* when using the 95% criterion (Table 3.4). The 0.01 and 50% criterion was little impacted by the value of *L* whereas the 95% criterion was substantially affected (Table 3.4).

Under both the 0.01 criterion and the 50% criterion, the number of OTUs is relatively stable with both showing similar counts, which is not surprising as the posterior density for the 0.01 criterion is generally at or near 50%. For the 95% criterion and decreasing values of *L*, winter decreases from 470 to 225 OTUs, spring decreases from 329 to 180 OTUs, summer decreases from 296 to 180 OTUs, and fall decreases from 424 to 222 OTUs.

It is noteworthy that value of *L* (pre-set maximum number of assemblages) did not have as dramatic effect on the number of predominant assemblages. There is some change when the number of assemblages is small, but in this case, when $L>10$ the results are relatively stable. Thus, my previous analyses using $L = 25$ assemblages are likely stable in terms of a well-estimated posterior assemblage distribution. Figure 3.7 (A and B) present the results graphically for the $L=5$ and $L=40$ scenarios; note that they seem quite similar.

Table 3.4
The effect of the number of assemblages (*L*) on the size (number of OTUs) in the predominate assemblage for the L4 samples.

Assemblages (*L*)	OTU criterion	No. OTUs in primary assemblage			
		Winter	Spring	Summer	Fall
40	95%	225	180	180	222
	50%	17	12	13	18
	0.01	17 (50%)	22 (62%)	17 (55%)	17(48%)
20	95%	250	201	209	245
	50%	18	13	13	18
	0.01	17 (49%)	22 (61%)	17(54%)	17(49%)
10	95%	304	245	227	288
	50%	19	13	14	19
	0.01	15 (46%)	21 (59%)	17(54%)	17(48%)
5	95%	470	329	296	424
	50%	20	14	14	20
	0.01	15 (45%)	21 (59%)	17(53%)	17 (47%)

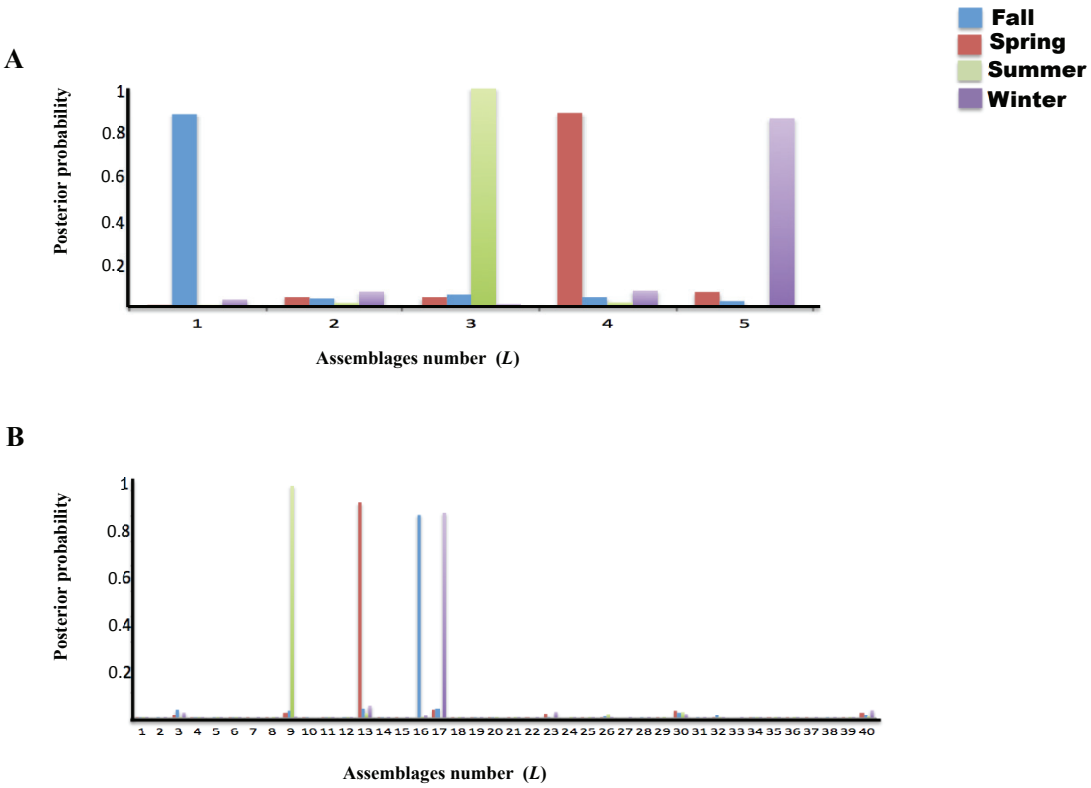

Figure 3.7 The assemblage distribution of L4 station from running different number of assemblages. (A) The assemblage distribution when L = 5. (B) The assemblage distribution when L= 40.

CHAPTER 4: CONCLUSIONS

In Chapter 2, the investigation of community structure at the global level showed that polar and tropical zones have distinct structures. However, the temperate zone did not have a distinct community structure, presumably because temperate samples varied greatly among location and seasons. In order to investigate this further, I carried out a series of analyses at a progressively more local scale (North Atlantic, L4 station and expanded English Channel). The L4 station of the English Channel represents an ideal reference site, as it is one of only a few North Atlantic temperate sites that have been extensively sampled for seasonal variation. A model re-trained on L4 equinox/solstice samples yielded very good predictions for the L4 site (and somewhat good prediction for nearby sites in the English Channel) but poor prediction for the other sites in the temperate North Atlantic. This result suggests that seasonal transitions among communities at a single Eastern Atlantic site (and possibly a localized region) are repeatable. Taken together with the results of El-Swais et al. (2015), who found predicable transitions between assemblage structures at a Western North Atlantic site, suggests that seasonal community assembly in the temperate North Atlantic is likely due to season-specific environmental drivers. The results also support the view of Becking (1934) that environmental factors are general drivers of microbial biogeographic patterns. The results are therefore consistent with the notion that the assembly of seasonal marine communities is based on growth of bacteria that are present at this location throughout the year, even if at times their abundances are very low (Gibbons et al. 2013).

The finding that the L4-based predictive model was somewhat predictive for nearby sites within the English Channel, but not as good as for the L4 site, suggests that

different temporal dynamics may be happening at those other sites. Presumably, all temperate sites go through similar transitions but at different times. However, the specific assemblages may also differ. Nonetheless, to the extent that the seed bank hypothesis is correct (Gibbons et al. 2013), and that seasonal transitions are predictable from year to year in the temperate North Atlantic, then the notion that there is a community-level response to environmental drivers may offer a general explanation for bacterial biogeographic patterns in the North Atlantic. This contrasts with the notion that biogeographic patterns for terrestrial zones are generally explained by endemism (Gibbons et al. 2013).

Chapter 3 presents four methodological investigations using the Polar and the L4 datasets. The first explored different ways of defining "predominant" OTUs by testing, >0.01 posterior probability (PP), 95% posterior density (PD), 50% PD, and inflection point criteria. The >0.01 PP was far more conservative (smaller number of OTUs), but only with the Antarctic/Arctic samples. For the L4 samples, the >0.01 PP criterion and the 50% PD criterion produced similar results. The 95% PD criterion was the most liberal. The inflection point analysis could not be used. The distinction between methods could have pragmatic significance. Clearly, the 95% PD criterion will always produce the largest number of OTUs, but too large a number may hinder subsequent analyses and interpretations. Conversely, the 50% PD and >0.01 PP criterion will likely find a lower number of OTUs, which would facilitate interpretation, but might miss some of the biodiversity. For example, when specifically comparing biodiversity among assemblages, one might use the 95% PD criterion. Conversely, when looking for "marker" species (informative OTUs), one might want the >0.01 PP criterion. The comparison of the 50% PD and >0.01 PP criterion alone has implications for biodiversity. If the >0.01 PP

criterion should explain a large proportion (*e.g.,* >50%) of the PD, then it is implies that much of the biodiversity is concentrated in a relatively few OTUs.

The user of BioMiCo has to make decisions about how the MCMC will be run, as well as the strategy for training and testing. The burn-in is important, but it seems that 1000 iterations should be sufficient. Also, the longer the run, the better the inference about the posterior distribution, and I recommend runs should be much longer than 500. Here, pragmatic considerations may be critical. Users will be limited by the speed of their computer and the size and availability of a cluster computer (which permits simultaneous analyses). A general recommendation is simply that users should run their MCMC as long as is practical. Also, the more samples used for training, the better the prediction. As a general recommendation, several replicates of 2/3 training and 1/3 testing should be adequate. Leave-one-out cross validation works well, but is computationally demanding, and is only necessary when datasets are so small that 2/3 training is infeasible. Lastly, results were not overly sensitive to the pre-specified number of assemblages (L) for the datasets used here. Although I expect that L=25 should be sufficient in many cases, I suggest that small-scale robustness analysis of alternative values of L could be carried out to identify if smaller values of L can be employed to save on the computational cost of the analysis.

In BioMiCo the OTUs contribute with the assemblages and the assemblages contributes with K factors (feature of interest). Each assemblage is comprised of a mixture of T different OTUs. The contribution of different OTUs to an assemblage is modeled by T mixing probabilities. In BioMiCo, a symmetric Dirichlet was used because there was no prior preference or knowledge about a particular assemblage structure. One question about the results obtained in this study is how the used of the Dirichlet prior

relates to over-fitting or under-fitting the inferred number of assemblages. The models were run with L =25, and the results obtained were consistently much less than 25. Every sample from a population will contain some statistical variation from the population points where the sample (seems to) deviate substantially from the population. This is just a simple consequence of sampling variation (random variation). If the model tries to address all of these "random fluctuations" it would be extremely over-fit. Our results ere far from this. Indeed the indication is that the model may tend to underfit (it has not captured all of the true signal in the data; i.e., the model seems to be too simple for the data). Specifically, the global analysis used only a single assemblage to describe the polar samples, yet subsequent analyses showed that there were distinct differences between the arctic and Antarctic communities. The tendency for BioMiCo to underfit the data seems to be due to using sparse and symmetric Dirichlet priors. Symmetry means that, initially, the prior will give no one assemblage an "advantage" in fitting the data. Thus, the assemblages are built from the data. However, by using a sparse prior the model will try to minimize variance, and this will likely lead to a tendency to use the smallest number of assemblages to fit the data . For this reason, it is not surprising that the model used one assemblage per factor, as shown in the polar case. Note that this is will not always be the outcome of model fitting (e.g., El-Swais et al. 2015)

In this study BioMiCo was used to predict global and local community structure of marina microbial community using 16S rRNA. PICRUSt could be used in future work to predict and explore functional profiles of microbial communities. Also, more comprehensive statistical investigation of how OTUs abundance at the L4 station is related to other biotic and abiotic factors is warranted. These analyses were not carried out as part of this study simply due to the lack of time. My results on the Eastern

Atlantic, when taken with the results of El-Swais et al. (2015) for the Western Atlantic, raises the question of whether similar environmental drivers may be operating across the North Atlantic. Greater temporal sampling (along with the collection of appropriate biotic and abiotic metadata) at a variety of sites across the North Atlantic is required before we can address this question. Given such data, the role of different environmental drivers of seasonal transitions could then be explicitly tested. To the extent that the global seed bank hypothesis is incorrect, and that dispersal via ocean currents also contributes to marine biogeographic patterns, the identification of such environmental drivers will be a daunting task.

BIBLIOGRAPHY

Armougom, F., & Raoult, D. (2012). Exploring microbial diversity using 16S rRNA high-throughput methods. *Journal of Computer Science & Systems Biology*, *2009*.

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

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

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Brown, E. A., Chain, F. J., Crease, T. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities?. *Ecology and evolution*, *5*(11), 2234-2251.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335-336.

Caporaso, J. G., Paszkiewicz, K., Field, D., Knight, R., & Gilbert, J. A. (2012). The Western English Channel contains a persistent microbial seed bank. *The ISME journal*, *6*(6), 1089-1093.

Carsey, T. P., Featherstone, C. M., Goodwin, K. D., Sinigalliano, C. D., Stamates, S. J., Zhang, J., . . . Adler, M. M. (2011). Boynton-delray coastal water quality monitoring program.

Chazdon, R. L., Colwell, R. K., Denslow, J. S., & Guariguata, M. R. (1998). *Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica* (No. Man and the Biosphere Series no. Vol. 20).

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., ... & Tiedje, J. M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research*, *37*(suppl 1), D141-D145.

Colwell, R. K. (2009). Biodiversity: Concepts, patterns, and measurement. *The Princeton Guide to Ecology, *, 257-263.

Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *345*(1311), 101-118.

Cottrell, M. T., & Kirchman, D. L. (2009). Photoheterotrophic microbes in the Arctic Ocean in summer and winter. *Applied and environmental microbiology*, *75*(15), 4958-4966

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.

Doney, S. C., Ruckelshaus, M., Duffy, J. E., Barry, J. P., Chan, F., English, C. A., . . . Knowlton, N. (2012). *Climate change impacts on marine ecosystems*. *Marine Science, 4*

Dove, A. (2013). Microbiomics: *The germ theory of everything*. *Science, 340*(6133), 763-765.

Dunn, K. A., Connors J. M., MacIntyre, B., Stadnyk, A., Thomas, N. A., Noble, A., Mahdi, G., Rashid, M., Otley, A. R., Bielawski, J. P., & Limbergen, J. V. (2016). Early changes in microbial community structure are associated with sustained remission following nutritional treatment of PediatricCrohn's Disease

El Swais, H., Dunn, K. A., Bielawski, J. P., Li, W. K., & Walsh, D. A. (2015). Seasonal *assemblages and short lived blooms in coastal north west atlantic ocean bacterioplankton*. *Environmental Microbiology, 17*(10), 3642-3661.

Engel, P., James, R. R., Koga, R., Kwong, W. K., McFrederick, Q. S., & Moran, N. A. (2013). Standard methods for research on Apis mellifera gut symbionts. *Journal of Apicultural Research*, *52*(4), 1-24.

Eren, A. M., Vineis, J. H., Morrison, H. G., & Sogin, M. L. (2013). A filtering method to generate high quality short reads using Illumina paired-end technology. *PloS one*, *8*(6), e66643.

Erlich, H. A., Gelfand, D., & Sninsky, J. J. (1991). Recent advances in the polymerase chain reaction. *Science*, *252*(5013), 1643-1651.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479-491.

Fuhrman, J. A., Cram, J. A., & Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, *13*(3), 133-146.

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., & Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences*, *105*(22), 7774-7778.

Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V., & Naeem, S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences*, *103*(35), 13104-13109.

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 Sciences of the United States of America, 109*(43), 17633-17638. doi:10.1073/pnas.1208160109 [doi]

Gibbons, S. M., Caporaso, J. G., Pirrung, M., Field, D., Knight, R., & Gilbert, J. A. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences of the United States of America, 110*(12), 4651-4655. doi:10.1073/pnas.1217767110 [doi]

Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., . . . Joint, I. (2009). The seasonal structure of microbial communities in the western english channel. *Environmental Microbiology, 11*(12), 3132-3139.

Galand, P. E., Potvin, M., Casamayor, E. O., & Lovejoy, C. (2010). Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *The ISME journal*, *4*(4), 564-576.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., & Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols, 2010*(1), pdb.prot5368. doi:10.1101/pdb.prot5368 [doi]

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews : MMBR, 68*(4), 669-685. doi:68/4/669 [pii]

Hansen, A. J., Neilson, R. P., Dale, V. H., Flather, C. H., Iverson, L. R., Currie, D. J., ... & Bartlein, P. J. (2001). Global change in forests: responses of species, communities, and biomes interactions between climate change and land use are projected to cause large shifts in biodiversity. *BioScience*, *51*(9), 765-779.

Heck, K. L., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, *56*(6), 1459-1461.

Hill, T. C., Walsh, K. A., Harris, J. A., & Moffett, B. F. (2003). Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology*, *43*(1), 1-11.

Hong, S. H., Bunge, J., Jeon, S. O., & Epstein, S. S. (2006). Predicting microbial species richness. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(1), 117-122.

Huber, J. A., Welch, D. B. M., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., & Sogin, M. L. (2007). Microbial population structures in the deep marine biosphere. *science*, *318*(5847), 97-100.

Hugerth, L. W., Muller, E. E., Hu, Y. O., Lebrun, L. A., Roume, H., Lundin, D., ... & Andersson, A. F. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One*, *9*(4), e95567.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and environmental microbiology*, *67*(10), 4399-4406.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., & Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*, *4*(11), e1000255.

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology, 45*(9), 2761-2764. doi:JCM.01228-07 [pii]

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, *35*(9), 3100-3108.

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., ... & Beiko, R. G. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology*, *31*(9), 814-821.

Levin, S. A., Carpenter, S. R., Godfray, H. C. J., Kinzig, A. P., Loreau, M., Losos, J. B., ... & Wilcove, D. S. (Eds.). (2009). *The Princeton guide to ecology*. Princeton University Press.

Lin, Y., Rajan, V., & Moret, B. M. (2012). A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1014-1022.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., ... & Tilman, D. (2001). Biodiversity and ecosystem functioning: current knowledge and future challenges. *science*, *294*(5543), 804-808.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, *71*(12), 8228-8235.

Lozupone, C., Hamady, M., & Knight, R. (2006). UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics, 7*(1), 1.

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, *73*(5), 1576-1585.

Lucas, J., Wichels, A., Teeling, H., Chafee, M., Scharfe, M., & Gerdts, G. (2015). Annual dynamics of North Sea bacterioplankton: seasonal variability superimposes short-term variation. *FEMS microbiology ecology*, *91*(9), fiv099.

Luria, C. M., Ducklow, H. W., & Amaral-Zettler, L. A. (2014). Marine bacterial, archaeal and eukaryotic diversity and community structure on the continental shelf of the western Antarctic Peninsula.

Martin, A. P. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and environmental microbiology*, *68*(8), 3673-3682.

Mount, D. W. (2008). Maximum parsimony method for phylogenetic prediction. *Cold Spring Harbor Protocols*, *2008*(4), pdb-top32.

Murray, A. E., & Grzymski, J. J. (2007). Diversity and genomics of Antarctic marine micro-organisms. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1488), 2259-2271.

Nishiguchi, M. K. (2000). Temperature affects species distribution in symbiotic populations of Vibrio spp. *Applied and Environmental Microbiology*, *66*(8), 3550-3555.

Patin, N. V., Kunin, V., Lidström, U., & Ashby, M. N. (2013). Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microbial ecology*, *65*(3), 709-719.

Prosser, J. I., Bohannan, B. J., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., ... & Osborn, A. M. (2007). The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, *5*(5), 384-392.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, 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.

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 [doi]

Robinson, C. J., Bohannan, B. J., & Young, V. B. (2010). From structure to function: The ecology of host-associated microbial communities. *Microbiology and Molecular Biology Reviews : MMBR, 74*(3), 453-476. doi:10.1128/MMBR.00014-10 [doi]

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., ... & Beeson, K. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*, *5*(3), e77.

Schloss, P. D., Larget, B. R., & Handelsman, J. (2004). Integration of microbial ecology and statistics: a test to compare gene libraries. *Applied and environmental microbiology*, *70*(9), 5485-5492.

Schloss, P. D., & Handelsman, J. (2006). Introducing TreeClimber, a test to compare microbial community structures. *Applied and environmental microbiology*, *72*(4), 2379-2384.

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 [doi]

Schuwirth, B. S., Borovinskaya, M. A., Hau, C. W., Zhang, W., Vila-Sanjurjo, A., Holton, J. M., & Cate, J. H. D. (2005). Structures of the bacterial ribosome at 3.5 Å resolution. *Science*, *310*(5749), 827-834.

Shade, A., Caporaso, J. G., Handelsman, J., Knight, R., & Fierer, N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *The ISME journal*, *7*(8), 1493-1506.

Shafiei, M., Dunn, K. A., Boon, E., MacDonald, S. M., Walsh, D. A., Gu, H., & Bielawski, J. P. (2015). BioMiCo: A supervised bayesian model for inference of microbial community structure. *Microbiome, 3*(1), 1.

Sheik, C. S., Jain, S., & Dick, G. J. (2014). Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. *Environmental microbiology*, *16*(1), 304-317.

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.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., ... & Cornejo-Castillo, F. M. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359.

Thomas, T., Rusch, D., DeMaere, M. Z., Yung, P. Y., Lewis, M., Halpern, A., . . . Kjelleberg, S. (2010). Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *The ISME Journal, 4*(12), 1557-1567.

Van de Peer, Y., Robbrecht, E., De Hoog, S., Caers, A., De Rijk, P., & De Wachter, R. (1999). Database on the structure of small subunit ribosomal RNA. *Nucleic acids research*, *27*(1), 179-183.

Wang, Y., & Qian, P. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PloS One, 4*(10), e7401.

Wang, Y., Tian, R. M., Gao, Z. M., Bougouffa, S., & Qian, P. Y. (2014). Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PloS one*, *9*(3), e90053.

Yang, F., Chia, N., White, B. A., & Schook, L. B. (2013). Compression-based distance (CBD): a simple, rapid, and accurate method for microbiota composition comparison. *BMC bioinformatics*, *14*(1), 1.

Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, *17*(1), 1.

Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B. M., Ramette, A. (2011). Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One, 6*(9), e24570.

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., ... & Scanlan, D. J. (2008). Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environmental microbiology*, *10*(1), 147-161.

Appendix A
Details about Sample Characteristics.

Table A1.
Sample source as classified by International Hydrological Organization (IHO Classification) for the ICOMM samples.

	Number	Percent
Arabian Sea	8	2.2
Arctic Ocean	17	4.8
Baltic Sea	8	2.2
Bay of Bengal	3	0.8
Beaufort Sea	3	0.8
Black sSea	5	1.4
Caribbean Sea	3	0.8
Coral Sea	16	4.5
Davis Strait	1	0.3
English Channel	68	19.1
Gulf of Aqaba	16	4.5
Gulf of California	2	0.6
Gulf of Mexico	3	0.8
Indian Ocean	1	0.3
Laptev Sea	1	0.3
North Atlantic Ocean	61	17.1
North Pacific Ocean	26	7.3
North Sea	32	9
Northwestern Passages	2	0.6
Norwegian Sea	1	0.3
Rio de la Plata	1	0.3
South Atlantic ocean	9	2.5
South Pacific ocean	36	10.1
Southern Ocean	20	5.6
Tasman sea	5	1.4
Coastal Waters: Southeast Alaska and British Columbia	8	2.2
Total	356	100

Table A2.
Sample Biome for the ICOMM samples

	Number	Percent
estuarine biome	2	0.6
mangrove biome	2	0.6
marine abyssal zone biome	17	4.8
marine basaltic hydrothermal vent biome	10	2.8
marine bathyal zone biome	32	9
marine benthic biome	2	0.6
marine cold seep biome	3	0.8
marine hydrothermal vent biome	22	6.2
marine neritic benthic zone biome	6	1.7
marine salt marsh biome	2	0.6
marine subtidal rocky reef biome	1	0.3
neritic epipelagic zone biome	156	43.8
neritic littoral zone	19	5.1
neritic mesopelagic zone biome	5	1.4
neritic sub-littoral zone	38	10.7
neritic supra-littoral zone	4	1.1
oceanic abyssopelagic zone biome	2	0.6
oceanic bathypelagic zone biome	2	0.6
oceanic epipelagic zone biome	29	8.1
oceanic mesopelagic zone biome	2	0.6
Total	356	100

Table A3.
Samples classified by habitat for the ICOMM samples

	Number	Percent
anoxic basin	4	1.1
brdu	4	1.1
coastal	221	62.1
microbial mat	2	0.6
open ocean	77	21.6
sediment	2	0.6
sponge	16	4.5
vents	30	8.4
Total	356	100

Table A4.
Samples classified by "features" for the ICOMM samples.

	Number	Percent
	41	11.5
abyssal plain	17	4.8
bay	1	0.3
cold seep	3	0.8
continental rise	10	2.8
continental shelf	13	3.7
cove	1	0.3
delta	1	0.3
estuarine bulk water	1	0.3
fjord	16	4.5
island	16	4.5
lagoon	2	0.6
mangrove swamp	2	0.6
marine anoxic zone	17	4.8
marine benthic feature	5	1.4
marine bulk water	17	4.7
marine hydrothermal vent	2	0.6
marine hydrothermal vent chimney	4	1.1
marine sponge reef	16	4.5
marine subtidal rocky reef	1	0.3
marine upwelling	14	3.9
marine wind mixed layer	112	31.5
mesoscale marine eddy	13	3.7
microbial mat	1	0.3
sandy beach	24	6.7
sea beach	6	1.7
Total	356	100

Table A5.
Samples classified by " type" for the ICOMM samples.

	Number	Percent
biofilm	10	2.8
Host_animal	12	3.4
sediment	102	28.7
water	232	65.2
Total	356	100

Table A6.
Samples classified by "material" for the ICOMM samples.

	Number	Percent
anaerobic sediment	13	3.7
anoxic water	20	5.6
biofilm material	4	1.1
brackish water	10	2.8
coastal water	119	32.4
contaminated sediment	2	0.6
fresh water	1	0.3
inorganically contaminated sediment	4	1.1
marine sediment	32	9
microbial mat material	10	2.8
ocean water	64	18
organic material	12	3.4
red clay	1	0.3
sandy sediment	25	7
sea water	16	4.5
sediment	3	0.8
silty sediment	16	4.5
suspended sediment	4	1.1
Total	356	100

Appendix B
Details about Individual Runs for the 3 Environment and 6 Environment Cases

Table B1
Predictions for each of the 5 Runs in the 3 Environment Case.

Run			Actual			
			Polar	Temperate	Tropical	Total
1	Predicted	Polar	**16**	23	1	40
		Temperate	0	**68**	1	69
		Tropical	0	11	**23**	34
		Total	16	102	25	143
2	Predicted	Polar	**16**	29	1	46
		Temperate	0	**56**	0	56
		Tropical	0	17	24	41
		Total	16	102	**25**	143
3						
	Predicted	Polar	**16**	23	4	43
		Temperate	0	**69**	0	69
		Tropical	0	7	**21**	28
		Total	16	102	25	143
4	Predicted	Polar	**16**	20	3	39
		Temperate	0	**73**	2	75
		Tropical	0	9	**20**	29
		Total	16	102	25	143
5	Predicted	Polar	**16**	29	1	46
		Temperate	0	**65**	2	67
		Tropical	0	8	**22**	30
		Total	16	102	25	143

Table B2
Predictions for each of the 5 Runs in the 6 Environment Case.

Run			Actual						
			Polar	Winter	Spring	Fall	Summer	Tropical	Total
1	Predicted	Polar	16	0	0	0	1	1	18
		Winter	0	9	23	16	13	11	72
		Spring	0	0	15	1	3	0	19
		Fall	0	0	2	8	2	1	13
		Summer	0	0	0	0	9	0	9
		Tropical	0	0	0	1	1	12	14
		Total	16	9	40	26	29	25	145
2	Predicted	Polar	13	0	1	0	0	1	15
		Winter	3	8	35	18	19	11	94
		Spring	0	0	3	0	0	0	3
		Fall	0	0	0	6	0	0	6
		Summer	0	0	0	1	8	0	9
		Tropical	0	1	1	1	2	13	18
		Total	16	9	40	26	29	25	145
3	Predicted	Polar	14	0	0	0	0	0	14
		Winter	2	9	31	19	20	17	98
		Spring	0	0	7	0	0	0	7
		Fall	0	0	1	7	0	0	8
		Summer	0	0	0	0	9	0	9
		Tropical	0	0	1	0	0	8	9
		Total	16	9	40	26	29	25	145
4	Predicted	Polar	15	0	1	0	0	0	16
		Winter	1	9	25	20	18	16	89
		Spring	0	0	13	2	3	0	18
		Fall	0	0	0	2	0	1	3
		Summer	0	0	1	1	8	0	10
		Tropical	0	0	0	1	0	8	9
		Total	16	9	40	26	29	25	145
5	Predicted	Polar	13	0	1	1	0	0	15
		Winter	3	9	30	21	19	16	98
		Spring	0	0	6	1	3	0	10
		Fall	0	0	1	3	0	1	5
		Summer	0	0	1	0	6	1	8
		Tropical	0	0	1	0	0	7	8
		Total	16	9	40	26	29	25	145

Appendix C
OUT ID and taxonomy for .01 criterion for 3 zones

#OTU ID	Taxonomy
5901	Root; k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__
8407	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria
84240	Root; k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__
95477	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria
105551	Root; k__Bacteria; p__Firmicutes; c__Bacilli
105774	Unassignable
151578	Root; k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria
158847	Root; k__Bacteria
234682	Root; k__Bacteria
277633	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria
306657	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria
317182	Root; k__Bacteria; p__Cyanobacteria; c__Chloroplast; o__Stramenopiles; f__
317708	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Oceanospirillales
319540	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Oceanospirillales; f__Halomonadaceae
511577	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria
528099	Root; k__Bacteria; p__SAR406; c__AB16; o__ZA3648c; f__AEGEAN_185
534609	Root; k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria
543795	Unassignable
557211	Root; k__Bacteria
New.0.CleanUp.ReferenceOTU103333	Root; k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__
New.0.CleanUp.ReferenceOTU105656	Root; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria

New.0.CleanUp.ReferenceOTU106808	Root; k__Bacteria
New.0.CleanUp.ReferenceOTU62958	Root; k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria
New.0.CleanUp.ReferenceOTU72937	Root; k__Bacteria
New.0.CleanUp.ReferenceOTU91013	Root; k__Bacteria
New.0.ReferenceOTU377	Root; k__Bacteria; p__Proteobacteria; c__Betaproteobacteria
New.0.ReferenceOTU398	Root; k__Bacteria
New.1.ReferenceOTU1050	Root; k__Bacteria
New.1.ReferenceOTU1479	Root; k__Bacteria
New.1.ReferenceOTU1750	Root; k__Bacteria; p__SAR406; c__AB16; o__Arctic96B-7
New.1.ReferenceOTU1790	Root; k__Bacteria
New.1.ReferenceOTU1798	Root; k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Erythrobacteraceae
New.1.ReferenceOTU2129	Root; k__Bacteria; p__Firmicutes
New.1.ReferenceOTU322	Root; k__Bacteria
New.1.ReferenceOTU390	Root; k__Bacteria
New.1.ReferenceOTU592	Root; k__Bacteria
New.1.ReferenceOTU769	Root; k__Bacteria; p__SAR406; c__AB16

Appendix D
OTU ID and taxonomy for Arctic and Antarctic (the Top 142)