ANALYSIS OF AQUATIC MICROBIAL COMMUNITIES IN EUTROPHIC AND HUMIC LAKES

by

Benjamin Crary

A thesis submitted in partial fulfillment of

The requirements for the degree of

Master of Science

(Environmental Engineering)

at the

UNIVERISTY OF WISCONSIN – MADISON

2013

Abstract

In this study, the microbial community structure and function of aquatic ecosystems were investigated. Illumina 16S gene tag sequencing was performed to profile the bacterial community of 7 humic lakes across multiple years of sampling, and shotgun metagenomic sequencing was performed to investigate the functional characteristics of humic and eutrophic lakes.

Lakes are significant in the global carbon cycle. In particular, humic bog lakes are reservoirs for large quantities of terrestrially derived organic matter, and thus link the aquatic and terrestrial carbon cycles. The bacterial communities that inhabit these ecosystems are of great interest to researchers, as they are the primary respirators of detritus. While DOC quality can be a strong driver of bacterial communities within lakes, resource availability varies greatly with space and time.

Here we sequenced the16S rRNA gene from over 700 samples from seven characteristic humic bog lakes to investigate the dynamics of bacterial community composition with time. Samples were taken weekly from each of the seven humic lakes during the ice-off season. For some lakes, communities from as many as four years were represented. This unique dataset allowed us to investigate the variability in diversity across temporal scales of years, seasons, weeks, and days, while maintaining the lens of spatial variation.

The diversity of each lake could be characterized adequately after approximately 25 samples that were spaced evenly throughout the ice-off season. Unpredictable year-to-year variation of diversity was observed, although, the normal range of Shannon Diversity was near 4. Diverse spring and early summer communities have been observed for aquatic bacteria in prior studies, but here we identified that on average, the community is also changing at the highest rate during early summer weeks in these lakes. Mixing events, which alter the physical and chemical factors of the habitat, also correlated with times of rapid change of community structure. Conversely, the community changed at the slowest rate during the strongly stratified periods of mid-late summer.

Metagenomic sequencing on two samples from one humic lake and three samples from a eutrophic lake revealed that the lakes were functionally most similar with space than season. The humic lake, which has a high concentration of humic substances from OM that is leached through a *sphagnum* mat, contained an overrepresentation of enzymes capable of degrading small organic acids and aromatic compounds. Presumably, photodegradation of humic substances are contributing the labile DOC pool for bacteria, and this represents a model for respiration of terrestrially derived C within aquatic ecosystems. Alternatively, the bacterial community in the eutrophic lake seemed suited for breaking down polysaccharides derived from algal biomass. Though more glycoside hydrolases were observed in the humic lake, the eutrophic community maintained a higher proportion of sulfatase enzymes that would putatively be active in degrading algal cell walls.

With these studies, we have identified that the diversity and change of bacterial communities in aquatic ecosystems vary multiple temporal scales, but the trends are generally shared for similar lakes. We also have seen that the functional potential of the bacterial community is more stable with time than across trophic levels, and the differences between the humic and eutrophic lake are largely a factor of substrate preference.

Acknowledgements

I would like to express my appreciation for the collaborators who have made this thesis possible. The commitments of individuals other than myself have greatly increased the quality of this work.

First, I would like to acknowledge the current and former members of the Microbial Observatory. Without their devoted sampling efforts, this thesis would be a completely different entity. These individuals include Sara Paver and Ashley Shade.

The staff at the University of Wisconsin's Trout Lake Station has also done a tremendous job of synchronizing the many research priorities that are active during the field season, which has allowed the Microbial Observatory to thrive for over ten years. I would like to extend my gratitude for those individuals who have contributed to this success, particularly Dr. Timothy Kratz and Pam Fashingbauer. Their efforts and cooperation are greatly appreciated.

I would like to thank the McMahon lab group for providing meaningful feedback and helping to move our collaborations forward. I have had the opportunity to work with Emily Read, Lucas Beversdorf, Ben Oyserman, Trevor Ghylin, James Mutschler, Robin Rohwer, Sarah Stevens, Travis Korosh, Pamela Camejo, and Pancho Moya. I am also grateful for the enthusiasm and quality of work I received from the two undergraduate students that I had the opportunity to mentor, Kelsey Phillips and Matt Bodin.

Outside of the McMahon lab, I want to thank Colin Fitzgerald and John Crawford for providing assistance and feedback on a variety of topics. I extend my gratitude to the lab of Dr. Paige Novak at the University of Minnesota-Twin Cites. In particular, Mark Krzmarzick and Patrick McNamara for mentorship and for introducing me to the field of research within environmental engineering.

I am thrilled to have had to opportunity to take courses from Dr. Daniel Noguera and Dr. Greg Harrington. Their commitment to the curriculum has made this experience challenging and rewarding. I would also like to thank Dr. Daniel Noguera and Dr. Garret Suen for serving on my M.S. committee.

Thank you to my family and friends, who have provided the support and motivation needed to continue through the difficult times . Particularly, my parents Bryan and Debra Crary, have given me the love and encouragement I have needed to reach my educational goals.

Finally, I would like to thank my advisor, Dr. Katherine McMahon. I appreciate the support she has provided throughout this process, and I am grateful to have had the opportunity to be guided by an individual who is enthusiastic about her work and the success of her students.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to Methods and Ecosystem

Historical techniques used to study aquatic microbiology and ecology

The whole of aquatic prokaryotic activity can perhaps be best described by the 'microbial loop' concept developed in 1983 (Azam et al. 1983), which enhanced an earlier theory of microbial food webs proposed nearly a decade earlier (Pomeroy 1974). The 'microbial loop' put forth by Azam and his colleagues described the role of bacteria in terms of autochthonous C utilization and energy production, stating that heterotrophic bacteria consume 10-50% of photosynthesized C and that they transfer that energy to higher trophic levels while being preyed upon by heterotrophic flagellates. While bacteria have traditionally been indicted in the role of nutrient and carbon mineralization on a global scale (Vernadsky 1945), Azam et al. took the first step to provide a pseudo-quantitative analysis of this idea in a marine ecosystem.

Historical references highlight that although the relative importance of bacteria had been assumed, there was no obvious means of quantifying their impact on an ecosystem during the middle of the 20th century. In lacustrine (Lindeman 1942) and marine (Riley 1951) ecosystems, researchers realized the role that heterotrophic bacteria must play in the flux of carbon and major geochemical cycles, yet did not have the molecular tools to support their hypotheses with quantitative evidence. With the development of staining techniques using the fluorescent dyes, Acridine Orange (Francisco, Mah, and Rabin 1973), DAPI (Porter and Feig 1980), and Hoechst 33825 (Paul 1982), researchers could begin to assess the abundance and biomass of aquatic bacteria. Due to the varying interactions of these dyes with cellular walls, moderate information on cell morphology could also be investigated with staining techniques. Concurrently, methods for overall bacterial production were also being developed, and early adopters of these methods were able to investigate the relationships of bacterioplankton and phytoplankton (Fuhrman, Ammerman, and Azam 1980; Fuhrman and Azam 1982) among other topics.

Though these early staining techniques provided a big leap forward in identifying patterns of overall bacterial communities, the tendencies and functions of populations, or even individuals, were still unknown.

The innovative work of Carl Woese and his colleagues in the 1970s and 1980s laid a foundation for our understanding of prokaryotic evolutionary diversity (Woese 1987). Over his career, Woese characterized the diversity within 16S and 18S small subunit rRNA gene (16S; 18S) sequences and established them as a means of quantifying the divergence of organisms via phylogenetic trees. Differences were found within cellular life, and the Eucarya, Bacteria, and Archaea domains were defined. Additionally, eleven bacterial phyla were described using 16S sequence similarities obtained from previously cultivated organisms. Subsequent efforts proposed the classic taxonomic organization: Phyla, Class, Order, Family, Genus, and Species (Holt and Sneath 2005). The foundation of this classification system is based on the conservation of the 16S gene. This approximately 1550 base pair long gene is present and conserved within all prokaryotes, but there are 9 ‘hypervariable’ regions in which the nucleotide sequence diverges (Figure 1-1). The nucleotide sequence within these hypervariable regions is still highly conserved within species, thus those regions can be targeted to assign taxonomy to an individual sequence or phylotype.

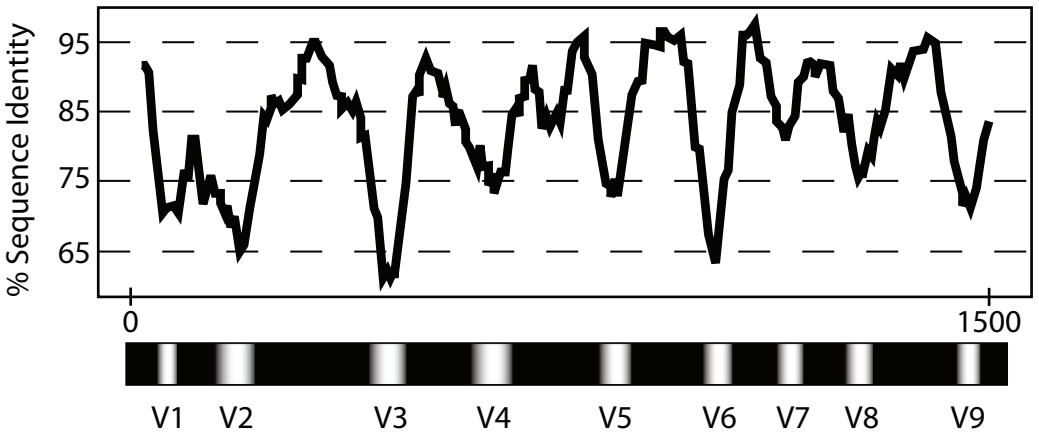

Figure 1-1: Hypervariable regions of the 16S small subunit rRNA gene

Another invaluable leap forward in all of microbial ecology was the development of polymerized chain reaction (PCR) to amplify DNA *in vitro* (Mullis and Faloona 1987). Using specific oligomer primers that can target desired genes, the relative quantification of specific functional genes or taxonomic marker genes (such as the 16S) could be achieved. PCR, which initially took advantage of a DNA polymerase enzyme from a thermophilic bacterium (*Thermus aquiticus*) before more efficient enzymes were identified, creates a reaction out of a polymerase enzyme, oligomer primers, and deoxynucleotides (dNTPs). During a series of temperature fluctuations that allows DNA to denature, primers to anneal, and polymerization at the primer binding sites, copies of a target gene were amplified. After approximately 40 repeated cycles, any gene or sequence that was originally in the sample should have been amplified if the oligomer primer successfully annealed to its target. Thus, cultivation became unnecessary to identify the bacterial populations present in an environmental sample.

While citations of discoveries based on PCR would be a complete review within itself, the value of PCR goes beyond any specific finding. Many molecular techniques have been made possible due to this technique. Phylogenetics based on taxonomic marker genes could be performed efficiently, whole community assessments could be performed with relative ease, and real-time quantification of expressed genes would eventually use the PCR approach. Additionally, early sequencing technologies were able to function due to the high quantities of target sequences generated, and genomes of isolated organisms could be captured.

Cultivation-independent techniques

Many efforts have been made to isolate or culture bacteria from freshwater, marine systems, soils, and other exotic environments. The attempts to isolate bacterial species from freshwater environments have yielded much knowledge on the growth patterns of bacterial species, but isolation experiments do not always prove to be a fruitful endeavor. That is, while some strains can grow successfully in isolation, other strains do not always survive without other community members. For example, there have been many strains isolated within some common freshwater genre, such as *Polynucleobacter* (Hahn et al. 2012; Hahn et al. 2009); however, pure strain of other ubiquitous genre, such as the acI of Actinobacteria have not been obtained. Still, the difficulties in acquiring isolates from these genres have yielded insights as to their lifestyles in a natural ecosystem. That is, when an organism cannot be isolated by selecting for its known niches, it indicates that there may be unknown interspecies interactions occurring. Growth experiments on isolates can also yield valuable ecological information, such as substrate preferences and growth rates.

Alternatively, cultivation-independent techniques can provide researchers with an idea of the community composition of a sample. While experiments cannot be performed on single populations, the response of whole communities or specific populations to experimental variables can be tracked. Amidst the development of such techniques, it had been assessed that less than 20% of all bacteria were "known", and thus methods that could provide could evaluate diversity were of great value (Wayne et al. 1987). Several similar techniques were developed and widely used, with some targeting the conserved ribosomal rRNA gene regions.

Automated Ribosomal Intergenic Spacer Analysis (ARISA) uses enzyme digestion sites, which are conserved among a population, on the intergenic region between the 16S rRNA and the 23S rRNA genes. After enzyme digestion, the remaining fragment lengths are used as a surrogate for unique individuals. ARISA is pseudo-quantitative (abundances are relative), which makes it possible to identify

the diversity and similarity within and among ecosystems. There is a history of aquatic microbiologists utilizing ARISA in an effort to identify the similarity of whole bacterial communities across lakes and across time, and their findings have helped confirm the repeating seasonal patterns of aquatic communities without the need for identifying the individual organisms (Fisher and Triplett 1999; Angela D Kent et al. 2007).

Similarly, Terminal Restriction Length Fragment Polymorphism (T-RFLP) (Liu et al. 1997) is a community fingerprinting approach that uses enzyme digestion, but on the 16S gene rather than on intergenic space, like is done with ARISA. The benefit of using the 16S gene is that it is possible to eventually sequence desired fragments from the digestion in an attempt to identify the phylogeny of the fragment. Ecologists have used this technique to understand community wide and taxon specific response to environmental variables. Applied to freshwater systems, ecologists have investigate the role of trophic status, chemical gradients, and the role of phytoplankton in structuring bacterial communities (Krzmarzick et al. 2013; Eiler and Bertilsson 2007; Paver and Kent 2010).

A third comparable whole community fingerprinting technique is denaturing gradient gel electrophoresis (DGGE). Though the seminal use of this method was for detection of single base-pair differences of nucleic acid sequences (Fischer and Lerman 1983), the method was later adopted to profile microbial communities with 16S gene amplicons (Muyzer, De Waal, and Uitterlinden 1993). In principle, 16S rDNA amplicons are denatured, via temperature or chemical means, and separated with gel electrophoresis. Fragment denaturation is dependent upon the nucleic acid sequence, and thus different genes in a sample will separate out during electrophoresis. By comparing two or more samples side by side, the difference in abundance of particular fragments (which are inferred as populations) is easily discerned.

Other cultivation-independent techniques have begun to produce species-specific knowledge that can be extrapolated onto the ecosystem level. One method, Fluorescent in situ Hybridization (FISH),

adapted the earlier bacterial staining techniques to yield phylogenetic information (Delong, Wickham, and Pace 1989). With FISH, strain-specific fluorescent oligomer probes (usually 16S SSU rDNA) can be used to identify which bacterial populations are within a sample in order to study community composition. This method is also quantitative, which is an advantage over the relative nature of ARISA, T-RFLP, and DGGE. Other techniques can be combined with FISH, such as microautoradiography (MAR), catalyzed reporter deposition (CARD) (Kerstens, Poddighe, and Hanselaar 1995; Pernthaler, Pernthaler, and Amann 2002), and electron labeling (EL) (Behrens et al. 2008) to link community structure with community function.

Culture-independent techniques can provide researchers with invaluable knowledge regarding which bacteria are active in metabolizing specific substrates in the environment. This information can be used to infer specific ecosystem functions performed by different species or the temporal dynamics of community members, but these techniques do not address evolutionary questions. Each of the described methods has their own limitations. ARISA, T-RFLP, and DGGE do not provide raw abundances, and diversity of a sample is inadequately covered. Variations of FISH are very powerful addressing the abundances or actions of several taxa at a time in a quantitative fashion, but this method cannot be used to address more than several populations at a time, nor does it address the interpopulation diversity. Recent sequencing technologies are able to overcome the issues of sampled diversity and interpopulation diversity, in an affordable and high-throughput manner.

Sequencing Technologies and Common Platforms

The availability of affordable genomic technologies has launched a period of rapid growth of knowledge surrounding the role of specific aquatic bacterial groups. The aforementioned "pregenomic" (referring to a period prior to high-throughput sequencing technologies) insights had begun to identify patterns among communities and populations, but sequencing technology has enabled researches to quantitatively compare individuals and populations at depths unobtainable by other techniques. Sanger sequencing (Sanger, Nicklen, and Coulson 1977) has been the most widely used sequencing technology since its development. Sanger sequencing is extremely accurate and has been employed for approximately 30 years, but it is limited by its cloning bias and relatively small throughput.

Methodologically, Sanger sequencing uses a fluorescently labeled oligomer primer to pair to the primer end of the fosmid or vector, and polymerizes the single strand of DNA with dideoxynucleotides (ddNTPs). Each reaction is run in four wells, with each well containing all four unlabeled deoxynucleotides (dNTPs) and a specific ddNTP. The ddNTP will cease polymerization because it lacks a 3'-OH group that is necessary for the phosphoester bond between nucleotides, thus the chain will terminate when a ddNTP is incorporated. When all four reactions are compared side by side with gel electrophoresis, one can determine the sequence of amplicon by observing the lengths of the terminated polymerized chains due to the labeled primer in each of the four lanes. Alternatively, chain termination variations use ddNTPs that are labeled with a dye unique to each base, which limits the necessary number of reactions to one. Modern chromatograms are typically employed after capillary electrophoresis as opposed to gel imaging.

Although the sequencing runs are extremely accurate, cloning biases plagues Sanger sequencing. The cloning bias arises from the necessity to insert the PCR amplicons into a plasmid vector to purify the PCR reaction. Some genes are unable to be inserted into a vector, because of toxicity to the vector itself (Sorek et al. 2007). Still, cloning bias can be minimized by the use of several types of vectors, and the

first sequenced bacterial genome was performed with Sanger sequencing on an isolate culture of *Haemophilus influenze* (Fleischmann et al. 1995). Many other genomes were also sequenced with this method. Further, sequencing of the 16S rRNA gene, much with Sanger sequencing, has elucidated the evolutionary structure of the tree of life.

Termed "next-gen sequencing technologies", methods such as pyrosequencing and sequencing by synthesis (SBS) have created an economical platform in which a researcher can acquire an enormous number of reads per run compared to traditional Sanger sequencing, while eliminating the need for vector cloning (Table 1-1). Pyrosequencing and SBS have accuracy and sequencing depth trade offs, but both are vast improvements over Sanger sequencing for those interested in studying whole communities. Though used for very similar purposes, there are clear methodological differences.

Pyrosequencing, which is often referred to as 454 due to the terminology used by the company who holds the patent, Roche (Margulies et al. 2005), is an extremely accurate method of sequencing that yields adequate amounts of sequence reads to characterize a bacterial community. In pyrosequencing, single-stranded DNA is bound to a primer before being mixed with polymerization enzyme, adenosine 5'-phosphosulfate (APS), and luciferin. Next, single nucleotides are added into the reaction and the nucleotide is synthesized into the complimentary strand if it is a match. Upon the incorporation of the nucleotide, pyrophosphate (PPi) is released and converted into ATP via sulfurylase enzymes (added to be present in the reaction mix). Luciferase subsequently reacts with the newly produced ATP, creating visible light that can be captured by a light sensitive device. This cycle is repeated until excess enzymes and byproducts the reactions inhibit further polymerization (Ronaghi, Uhlen, and Nyren 2013).

The major idea behind SBS is the addition of fluorescently labeled nucleotides. To start, a sample of double stranded DNA fragments are ligated to an adapter oligomer. DNA fragments loaded onto a flow cell that contains a dense surface of oligomer primers. The primers on the flow cell hybridize isothermally with the adapter on the single strand of DNA, and a subsequent incubation with isothermal polymerase

results in multiple amplifications of each sequence into a cluster on the flow cell. Once clusters are formed, reverse strands are cleaved and removed from the cluster, and random primers are added to the flow cell and bound to the ends of the single stranded clusters. Next, fluorescently labeled and reversibly terminated nucleotides are added in rounds. After a nucleotide is synthesized to the fragment, a laser excites the cluster that results in the emission of a color corresponding to the fluorescent label on the newly synthesized nucleotide. A light sensitive device captures each round, and color patterns from each cluster are interpreted into sequences.

Table 1-1: Technical comparison of common sequencing platforms

	Roche 454	**Illumina (HiSeq 2000)**	**Sanger**	**Pacific Bio**	**Ion Torrent (PGM)**
Method	Pyrosequencing	Sequencing by synthesis (SBS)	Chain termination	Single-molecule-real-time sequencing	Proton detection
Read Length	~ 700 bp	50 – 250 bp	400-900 bp	Average 1500 bp	~ 200 bp
Bp per run	700 Mb	600 Gb	1.9 -84 Kb	100 Mb	20 Mb-1 Gb
Time/run	1 day	3-10 days	3 hours	2 hours	2 hours
DNA required	50-1000 ng	100-1000 ng	1000 ng	50-1000 ng	50-1000 ng

Pyrosequencing and SBS can be employed after targeting specific gene regions with PCR. For example, the Earth Microbiome Project (Gilbert et al. 2010), which aims to characterize bacterial communities and populations in as many environments as possible in a standard way, amplifies the V4 hypervariable region of the 16S rRNA gene prior to sequencing (Figure 1-1). After amplifying a specific target, the typical sequencing protocol takes place. If the sequences target, or "tag", the 16S gene, they are sometimes referred to as "pyrotags" (if performed with pyrosequencing), "itags" (if performed with Illumina SBS technology), or more generally "tag sequences" (DeAngelis et al. 2010; Degnan and Ochman 2011; Galand et al. 2009).

Alternatively, PCR can be eliminated and shotgun sequencing can be utilized. The method of shotgun sequencing simply refers to the shearing of all DNA recovered from an extraction into strands that are short enough to be sequenced by current technology (~1-3 kb) followed by sequencing on any platform. Shotgun sequencing has been used on isolate cultures, with single cells sorted by flow cytometry (Martinez-Garcia et al. 2011), and for whole community genome analysis, or metagenomics (Handelsman et al. 1998) on all sequencing platforms.

16S small subunit rRNA (16S) Tag Sequence Analysis Techniques and Software:

The goals for studies that use 16S tag sequencing can range greatly. These goals can be to detect a particular organism that is suspected of performing a function of interest *in situ* (Kostka et al. 2011), to understand the dynamics of population abundance with time (Eiler, et al. 2011), or to understand population diversity with space (Fierer et al. 2011). The goals can also be community centric, where the focus may be to assess the diversity of a community with time and space (Gonzalez et al. 2012; Ashley Shade et al. 2013), or to assess the response of communities to a disturbance (Huse et al. 2008). Regardless of the specific goals, the analysis of tag sequences follows a typical analysis workflow (Figure 1-2) based on clustering sequences into operational taxonomic units (OTUs). This is often performed as part of standardized automated pipelines (Schloss et al. 2009; Caporaso et al. 2010). Note that although there are standardized pipelines, there is much ambiguity in the criteria used in these pipelines that ultimately leads to variations in the end result.

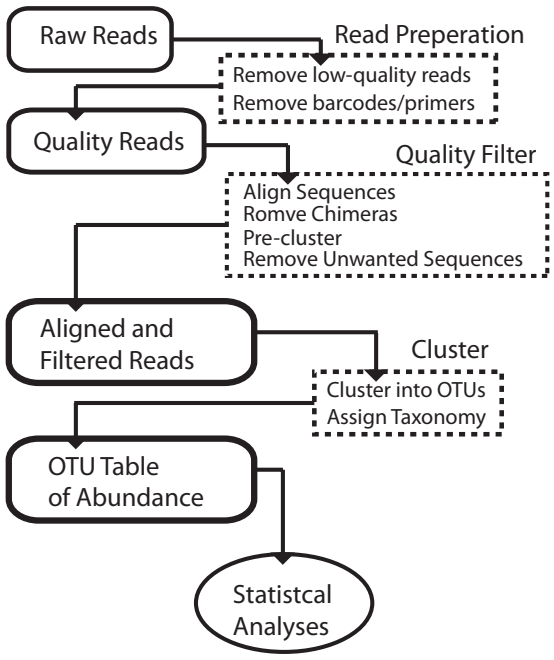

Figure 1-2: Steps included in a common sequence analysis for 16S rRNA gene reads

Before describing a generic analysis workflow, it is worth addressing what is often the end result of the pipeline: Operational Taxonomic Unit (OTU) abundance. The definition of an OTU is currently, and may always remain, a trending topic among ecologists, computational biologists, and bioinformaticians. Broadly, an OTU may be defined as a group or cluster of related reads that are assumed to be a coherent entity within the environmental sample from which it was sequenced. The debate begins when researchers try to define % sequence similarity within a cluster that is relevant to a 'species'. This debate is especially relevant for those that are attempting to assess diversity across samples. A rough standard has defined a species as a cluster of reads that share nucleotide 97% similarity across the full 16S gene, a genus as a cluster that shares at least 95% similarity, and a phylum as a cluster that shares at least 80% similarity (Schloss and Handelsman 2005). Thus, OTUs are clustered at 97% similarity if a species estimate is desired. As stated above, however, there is much debate about the environmental relevancy of these cutoffs (Stackebrandt and Goebel 1994; Patin et al. 2013; Sun et al. 2012).

OTU decisions will assuredly affect the overall results and conclusions of a study in at least a small way, but that decision is certainly not the only one that is made throughout a sequence analysis workflow. Before OTU definitions need to be considered, the initial step in a standard analysis workflow begins with the conversion of the raw sequence data into a readable nucleotide sequence. The data that is recovered from a sequencing run will vary upon the sequencing platform, but it usually contains some type of base call and a corresponding quality reading. The researcher must make appropriate decisions on what the minimum quality reading from the sequencing process should be considered acceptable for incorporation into read. The next step is to remove the primer sequence and unique barcode adapter sequence (if more than one sample are sequenced in a single run) that are carried over into the sequencing read. The primer can "waste" 12-18 base pairs of the available sequencing range, and the barcode may "waste" an additional 8 base pairs.

Once the sequences are called and trimmed of uninformative bases, an alignment to a known reference database, such as Silva (Pruesse et al. 2007), is a common approach. Alignments can be made with primary, secondary, or tertiary structures to the desired reference set with increasing computational requirements. Alternatively, the read set can be aligned *de novo*, without a reference set. The alignment step is necessary to ensure that the reads in fact are amplified of the region intended and to provide a means to quantify the similarity between read sequences.

After alignments are made and any sequences that do not align appropriately are removed (i.e.: any sequences that align out of the "standard" range from the rest of the reads), the read set is checked for chimeric sequences. Chimeric sequences are artifacts of inefficient PCR, and result in hybrid sequences that appear as novel OTUs (Haas et al. 2011). There are several programs that have been written to identify chimeric sequences within a dataset, each with computational or accuracy tradeoffs (Wright et al. 2012; Haas et al. 2011; Edgar et al. 2011). Chimera checking may be a final quality control step, but a preclustering step may be performed to lump reads that differ in as few as 1 bp per 100. This approach was developed to mask any errors in sequencing that may arise, in an effort to minimize the computational resources that would be needed to calculate the pairwise similarity between all unique sequences (Huse et al. 2010).

At this point, the read set is considered cleaned of any PCR or sequencing artifacts. The read set may still contain reads that originated from mitochondrial or chloroplasts that were present in the environmental sample, and these reads need to be removed from the dataset or they will skew the diversity and species results. To remove these, each sequence must be classified against a reference database. An increasingly common method for classifying sequences is the use of kmer comparison (Wang et al. 2007). Briefly, each query and reference sequence is checked for the frequency of kmers (each possible sequence of k nucleotides), and the query is assigned the taxonomy of the closest match. This method can be bootstrapped to provide a confidence score. In this analysis step, any unwanted assignments, such as chloroplast or mitochondria, can be removed. Once the read set is cleaned of quality

errors and uninformative reads, it is common practice to sub sample to a common number of reads if more than one sample is being processed.

Finally, the clustering of sequences into OTUs can be done. The two major competing approaches are used in Mothur and QIIME. The approach in Mothur is based upon creating a distance matrix for each pairwise comparison of sequences, and clustering sequences based these dissimilarities. The QIIME default uses the program Uclust, which attempts to pair each sequence to a reference dataset. While there is merit to both clustering approaches, it has been noted that the difference in accuracy between these two methods is minor compared to the variations in accuracy due to preprocessing steps. Although the default settings for Mothur and QIIME are considered here, but users of both Mothur and QIIME have the option to use several different algorithms.

Distance Based Clustering

As stated above, the Mothur workflow is based upon a distance matrix comparing all reads in the dataset. The pairwise similarity, or pairwise distance, between unique reads does involve a few subtleties. Gaps, mismatches, and ends can all be handled differently. First, gap penalties can be given for a single gap spanning multiple bases, or a gap for each missing base. The default in Mothur is a single gap penalty, due to the logic that a multiple base gap is likely due to a single insertion. Other programs such as DNADIST (Felsenstein and Churchill 1996) do not penalize for gaps, thus this program may be worth reconsidering if the read set is known or assumed to include many insertions. If using a BLAST (Altschup et al. 1990) based dissimilarity metric, mismatches and gaps can be weighted to adjust to the users preferences. Finally, the ends can be removed from or included in the distance calculation. The obvious concern with the read end is whether it represents a true insertion or if it simply was the end of the quality sequencing.

Following the creation of the distance matrix, a clustering algorithm is used to group reads into OTUs. Three algorithms are included in Mothur at present time; the nearest neighbor algorithm, the furthest neighbor algorithm, and the average neighbor algorithm. If a 97% OTU definition is chosen, the nearest neighbor will create clusters such that "Each of the sequences within an OTU is at most 3% distant from the most similar sequence in the OTU." The furthest neighbor algorithm creates clusters such that "Each of the sequences within an OTU is at most 3% distant from all of the other sequences within the OTU." The average neighbor algorithm takes the mean of the nearest and furthest neighbor algorithms (Schloss et al. 2009). While the furthest neighbor is default within Mothur, the average neighbor is slightly more accurate but more computationally expensive. For studies that have little use for OTU clusters beyond 97% similarity, Mothur has computational advantages, in that is allows the use to stop calculating the dissimilarity between two reads beyond a certain cutoff.

Library Based Clustering

The default clustering in QIIME uses the Uclust program (Edgar 2010). Clustering with Uclust can be completely *de novo* but a reference 'library' set can also be used. Each read is compared against sequence 'seeds' in the library. If the read does not cluster to a present seed, the read creates a new seed for which all other reads are compared against. The user predefines two sets of criteria before running the algorithm. The first is such that "all centroids have similarity < T to each other." The second is such that "all member sequences have similarity >= T to a centroid" (Edgar 2010). Caution should be used with this approach, as UCLUST is a greedy algorithm. This simply means that each read is run linearly through the reference library, and the algorithm will stop as soon as it finds a match to the criteria. This does not ensure that the match is optimal, and the order of the reference library needs to be adjusted to present an optimal order. While disadvantages exist, the Uclust algorithm does not use nearly as much

computational memory as the creation of a pairwise distance matrix. Thus, this method can be applied to deeper sequencing sets, or sets that include a large number of samples.

Alternative clustering approaches

One clustering alternative is oligotyping (available from http://oligotyping.org). Oligotyping is a computational method that detects particular nucleotides that have observable variation across query sequences from closely related taxa, which can allow the identification of distinct populations that may not be found with 3% clustering methods (Eren et al. 2011). The method is based upon finding the enthalpy of a particular nucleotide for an aligned query set. The enthalpy is based upon the percent of sequences that contain A, T, G, and C. In locations in which 100% of the sequences contain an 'A', then the enthalpy is very low because there is low variation at that location. For locations in which 50% of sequences have an 'A', 25% have a 'T', 12.5% have a 'C', and 12.5% have 'G', the enthalpy would be very high because there is much variation at that location. Over an entire query sequence length, the locations with high enthalpy are used to identify distinct 'oligotypes' or populations. The computational time is very short for the oligotyping process, but the necessary steps to acquire a query of aligned sequences from closely related organisms may still be computationally expensive.

Post clustering analysis

Taxonomic assignment becomes the final piece of a standard analysis workflow to produce results that can be interpreted through statistical means. As with most analysis steps, there are a number of options to assign taxonomy to an OTU. The Ribosomal Database Project (RDP) classifier has become widely used in both Mothur and QIIME (Wang et al. 2007). This classifier is a naïve Bayesian one, in the sense that each feature (in this case each base) is considered independent of the any other feature. This

approach does not require the alignments of sequences (although this is performed in most sequence analysis to construct OTUs) and is a rapid and accurate method of the assignment given full-length or partial sequences.

The algorithm at use is based on comparing kmer frequency between a reference taxonomy set and a representative read from each OTU. During method development, the accuracy of 6,7,8, and 9-mers were compared in their ability to accurately and confidently assign taxonomy. The 6 and 7-mers were relatively inaccurate, while 8 and 9-mers were comparably accurate. The 8-mer was chosen, as it is less computationally expensive to calculate. To provide a confidence for the assignment, one-eighth of the 8-character 'words' are sub-sampled 100 times and the number of times the same assignment was chosen is used as the percent confidence. This algorithm allows for hierarchical assignments to be made, since a bootstrapping will give confidence for difference taxonomic levels.

While the RDP classifier is widely used, there are other options. Aligning query reads to a reference database using BLAST is one such option (Altschup et al. 1990). Standard BLAST cutoffs, such as e-value, bit score, and percent id, can all be used as cutoffs to assign taxonomy. While perhaps very accurate, this method does not perform any hierarchical assignments. Other assignment alternatives include GAST (Huse et al. 2008) and SimRank (DeSantis et al. 2011).

The reference datasets themselves have a great impact upon the overall quality of results. A reference set with defined taxonomy must contain a sequence that is similar enough to a query sequence in order to assign proper taxonomy. Constructing such references can be a difficult task, especially if the rare biosphere is of interest. Several expansive reference databases exist, and these are frequently updated to include novel references.

Silva (Pruesse et al. 2007), RDP (Cole et al. 2009), and Greengenes (DeSantis et al. 2006) are the most commonly used alignment and taxonomy databases. Each of these databases has aligned their references sequences with secondary structure of the 16S gene, yet the alignment quality of each of these

databases varies. The hierarchical nomenclature proposed by *Bergey's Manual of Systematic Bacteriology* is the standard used by these three public reference databases. Studies have shown that the Silva database provides superior alignments to Greengenes over variable regions, while RDP does not align those regions (Schloss 2010). Upon the next update (release-115), the Silva database will contain just fewer than 4 million small subunit rRNA sequences that are at least 1200 bases in length. At the current time, RDP and Greengenes contain just over 1 million aligned 16S small subunit rRNA sequences that are at least 1200 bases in length, each.

Despite that Silva maintains the largest and most carefully aligned dataset, other considerations must be given before opting to use a particular database. For instance, it is of particular importance to make sure that the reference dataset is representative of the environment that is of question. Thus, custom aligned taxonomy databases can be used to predict taxonomies if public databases are not adequate at characterizing a particular environment (Newton et al. 2011).

Metagenomic Sequence Analysis Techniques and Software

Metagenomic sequencing technologies allow researchers to go 'beyond the genome' in their endeavors to answer questions of ecology, evolution, and engineering. Bacteria inhabit any environment capable of sustaining life, and are typically the sole denizens of extreme environments that will not support higher organisms. These microorganisms assume a hidden but essential role in biogeochemical cycling of almost all elements in their surrounding environments, and have thus played a critical role in shaping the earth as we know it. The importance of bacteria cannot be overlooked in any environment; bacteria are primary recyclers of nutrients and organic matter in aquatic ecosystems (Cotner and Biddanda 2002), soil bacteria fix nitrogen and provide nitrogen to crops during symbiosis (Barea et al. 2005), and cells in our own human bodies are outnumbered by bacterial cells (Berg 1996).

The study of cultured genomes has yielded many insights about the role of genomic structure in evolutionary strategy and ecological function, but the dependence upon cultivation had limited the field of genomics prior to high-throughput shotgun sequencing. The small percentage of organisms that can be isolated has biased our genomic databases in a way that is unlikely to be representative of the entire microbial community (Amann et al. 1995; Pace 1997; Rappé and Giovannoni 2003), and indeed, cultivation-independent sequencing studies have shown that our growing understanding of biodiversity is being matched by our growing understanding of functional diversity (D. Wu et al. 2009). The abundance of data from metagenomic sequencing is fast-paced and discovery driven, but until recently, population level information was difficult to interpret from metagenomes. Advances in computational methods have since allowed draft genomes to be acquired out of metagenomes based on coverage and kmer profiles, provided two or more samples are sequenced (Albertsen et al. 2013). To get to the point of removing draft genomes from samples, however, there are a number of significant computational steps that are taken once the sequence read data is recovered from the sequencing platform (Figure 1-3).

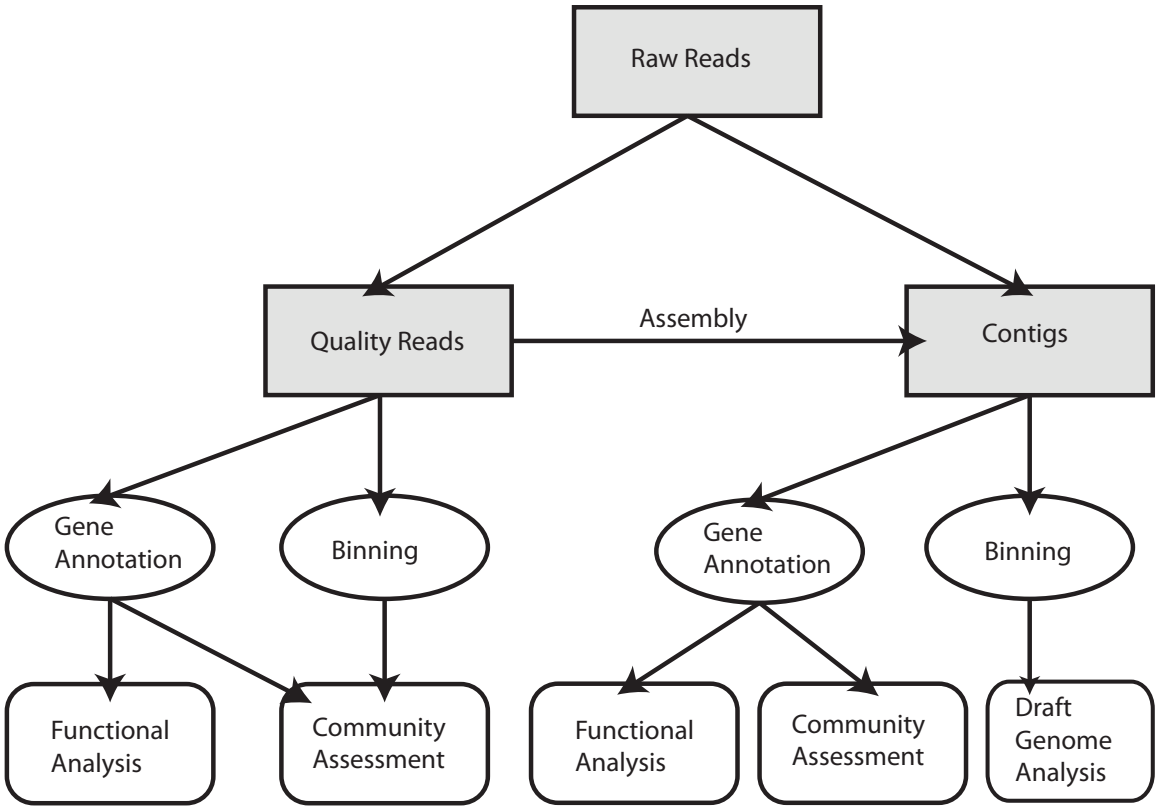

Figure 1-3: Typical workflow for metagenomic projects. Grey boxes represent sequences libraries; white circles and boxes represent analyses.

Preprocessing and Assembly

As with 16S tag sequencing, there are preprocessing steps and quality filters that must be in place upon getting read data from a sequencing run. First base calling is performed, most commonly with the program phred (Ewing et al. 1998). phred is a probabilistic-based base caller, in which raw data is converted into accuracy probability base calls. The nucleotide error rate varies depending upon the sequencing technology used, but error rates among called bases are less than 1% for Illumina HiSeq and Roche 454 high-throughput sequencers (Glenn 2011).

Once high quality reads are obtained, these reads are typically assembled into contiguous sequences. The process of assembly finds overlapping similarity among the reads and effectively 'stiches' them together to form consensus contigs. Repeating regions are likely to cause assembly errors (Lapidus

2009) and nonuniform species abundance has the potential to cause erroneous nonuniform read depth across in contigs coassembled from reads originating in distinct populations (Kunin et al. 2008). Due to the prevalence of chimeras in contigs that were shorter than 10 kbp, some have made the argument against read assembly (Mavromatis et al. 2007).

Nonetheless, the acquisition of contigs with long lengths can be very insightful due to the possible identification of functional operons. Phrap, Arachne (Batzogulou et al. 2002), JAZZ (Aparicio et al. 2002), Forge, and the Celebra assembler were all created for Sanger sequencing and long length reads. Upon the increased depth of short length reads from high-throughput sequencing, Velvet (Zerbino and Birney 2008), EULER (Pevzner, Tang, and Waterman 2001), Newbler (Roche/454 Life Sciences, Branford, CT, USA), SOAPdenovo (Li et al. 2010), and others have been created to assemble reads *de novo*. Contemporary genomic and metagenomic projects employ multiple assembly methods in order to converge on the most accurate contig sequence (Oh et al. 2011; Garcia et al. 2013; Shih et al. 2013). For in depth analysis of the algorithms used, refer to (Miller et al. 2010).

Gene Calling and Annotation

Most studies utilizing metagenomic sequencing are most interested in understanding the functional capabilities of the bacterial community. While reads may be in hand, there is much consideration that needs to be taken to predict genes encoded by the reads. Due to the short read and contig lengths and the overall complexity of metagenomic datasets, identifying open reading frames (ORFs), and subsequently genes, can become a difficult task. This task is made even more difficult in that the accuracy of gene prediction can be dependent upon preprocessing and assembly accuracy. Despite the challenges, gene prediction from simulated data could be predicted at 85% accuracy on assembled reads for highly complex communities, although estimates showed 70% accuracy with unassembled reads (Mavromatis et al. 2007).

Two approaches, “evidence-based” and “ab initio”, are generally used for gene calling (Kunin et al. 2008). The evidence-based approach translates nucleotide sequences into amino acids and searches against protein databases using BLAST. Caution must be given with this technique due to potential filtering of alignment matches with BLAST. Short read queries that originate from distant relatives of any organism in the reference set have the potential to be missed as genes. CRITICA (Badger and Olsen 1999) and ORPHEUS (Frishman et al. 1998) are two common programs that use this approach. Alternatively, the ab initio approach is based on the identification of protein coding and noncoding regions within the sequences, usually using monocodon frequency analysis (Azad and Borodovsky 2004). This approach benefits from being independent of a reference database for the identification of genes, thus novel reading frames can be identified. Programs such as Genemark (Besemer and Borodovsky 2005) and fgenesB (http://softberry.com]) use this technique. Drawbacks include the failure to identify partial ORFs which are true genes. Most often, gene calling is performed in pipelines that use multiple tools simultaneously.

Upon having called genes into ORFs, functional annotation is the next step in a typical metagenomic workflow. Functional prediction relies completely on curated reference databases, and novel genes without homologs in a curated database will yield little to no information about its enzymatic function. Thus metagenomic datasets from environments that are historically understudied may have limited success in functional prediction. That is not meant to dissuade future studies in rare environments, as these types of studies still have great value, particularly in enzyme discovery (Wilson and Piel 2013).

Many analysis tools provide biological inference of genes based on curated reference sets. These tools include SEED (Overbeek et al. 2005), KEGG (Kanehisa and Goto 2000), COG (Tatusov 1997), and EC (Bairoch 2000), among others. Generally, these tools compare nucleotide sequences to nucleotide and protein sequences of known protein families and pathways, so that inferences can be made on the query sequences. As noted by Chen and colleagues, the methods for which these references are predicted,

maintained, and compared against can vary greatly (Chen et al. 2013). Consequently, inferences about a metagenome may differ if more than a single database is referenced.

Analysis pipelines such as IMG-MER (Markowitz et al. 2012) and MG-RAST (Meyer et al. 2008) upload results form multiple functional annotations in an effort to settle the ambiguity between annotation methods. In particular, IMG-MER assigns function to protein coding regions using COG, Pfam, EC, and KEGG, so that comparisons can be made side by side. The benefit of using pipelines is that comparison across other public datasets is possible, and the analyses were made with the same criteria.

One common strategy to acquire the annotations is to use sequence similarity. Sequence similarity based annotations are made in IMG with the following methods: the COG functional annotations in IMG-MER are made using RSP-BLAST with a Position Specific Scoring Matrix, KEGG annotations are made using BLASTp; Pfam and TIGRfam annotations are made Hidden Markov Model searches with a BLAST prefilter (Mavromatis et al. 2009). Annotations made to COG or PFAM databases based solely on sequence similarity can yield insufficient information due to poorly characterized genes. Thus, additional methods utilizing orthology detection methods, analysis of synteny (Mavromatis et al. 2009), reconstruction of phylogenetic trees (Altenhoff and Dessimoz 2009), and specificity determining positions (Mazin et al. 2010) can be used with reference databases to provide refinement to non-specific functional annotations.

Community Assessment and Phylogenetic Binning

Assessment of the populations present in a metagenomic sample is not necessarily an easy task. The traditional 16S gene is not always a single copy within a genome, thus mapping metagenomic reads to this particular taxonomic marker will bias community composition towards genomes with multiple

copies. The same biases are present in tag sequencing, however, and the 16S reference taxonomies are by far the most robust. Other single copy alternatives have been suggested, such as rpoB and recA (Walsh et al. 2004).

Nonetheless, by aligning metagenomic reads to a taxonomic marker reference set with BLAST or other alignment tools, a researcher can retrieve sequences that yield information regarding community diversity and composition. Hao and Chen found that 150 bp long fragments of the 16S gene retrieved from metagenomic datasets would yield results nearly as accurately at full-length 16S sequences using a custom workflow (Hao and Chen 2012). In this method, reads were mapped back to a reference set with known taxonomy.

Binning all metagenomic reads to interpret community composition is an alternative to retrieving taxonomic marker genes from a metagenomic dataset. Assessment of species-specific functional attributes will also be possible with this approach. Several techniques are commonly used to bin metagenomic reads. Some of these are composition based (reference-free), while others are similarity based (reference-dependent).

Several common reference-free binning techniques exist. Binning based on GC content is on way of differentiating sequence origin, but finer resolution is available via the use of kmer frequencies. Similar to the approach used in the taxonomic classifications in Mothur, programs such as TETRA take advantage of the powerful statistical analysis to distinguish the species level origin of a sequence (Schbath, et al. 1995; Teeling et al. 2004). In TETRA, briefly, the expected frequency of each of the 256 combinations of tetranucleotides is estimated by a maximal-order Markov model, and the representation of each tetranucleotide is expressed as a Z-score.

While kmer binning is not error free, it has a large advantage in that it does not require a reference database to bin sequences. Similarly, CLaMs is another program that does not rely on references to bin reads, but can taxonomy can be placed to these bins if references are available (Pati et al.

2011). This program identifies genomic signatures derived from de Bruijn chains (Heath and Pati 2007). Codon usage is another way to bin metagenomic reads or contigs. Species use codons in differential frequencies to code for the same amino acids (an evolutionary tool to quickly ramp up protein production), and this knowledge can be applied to distinguishing between similar sequences in order to bin them based on origin.

Alternatively, similarity based binning is typically carried out with BLAST or with trees with a reference sequence set. Programs such as MEGAN (Huson et al. 2007) can take these results and assign each query sequence to its lowest common ancestor on the tree for BLASTs that have been performed against subjects that contain phylogenetic tree information (i.e.: NCBI nr). CARMA (Gerlach et al. 2009) is alternative program that uses Pfams for taxonomic classification. Although this approach is powerful for metagenomes that are low complexity and from well-characterized environments, a high number of genes with no homologs in reference databases can limit the applicability of similarity based binning.

Retrieval of Draft Genomes from Metagenomes

Two recent studies have demonstrated the ability to reconstruct nearly complete draft genomes of rare organisms from metagenomic data (Wrighton et al. 2012; Albertsen et al. 2013). The premise is based on having a collection of samples (at least two) with similar communities, and plotting the contigs that occur in each sample on emergent "self-organizing" maps. The contigs from distinct organisms separate out on these maps by forming clusters with the same tetranucleotide frequencies, abundance patterns, and GC %. A second assembly on the contigs that form these distinct clusters will yield a reconstructed draft genome, and completeness can be measured by the presence of conserved gene sequences. Wrighton et al. estimated greater than 90% completeness for 21 of 40 reconstructed genomes, even for organisms that made up less than 1% of the community (Wrighton et al. 2012). Comparable results were observed for the method defined by Albertsen et al. (Albertsen et al. 2013).

The totality of methods applied to metagenomics was certainly not covered here. As advances are made in the field of computer science, novel tools for metagenomes will likely follow. The methods outlined here (Figure 1-3) are generally representative of the basic approaches taken to make ecological or evolutionary inferences out of metagenomic data. Specific statistical tests will be applied to individual studies, but most often, metagenomic studies will need to address at least a portion of these sequence analysis steps.

Lake Characteristics

Limnology, the study of inland lakes, rivers, and wetlands, has long roots as a multidisciplinary science. Applied chemistry, physics, geology, and engineering have all left their mark on the filed. The result is a growing characterization of the physical, chemical, and biological elements of lakes. Lakes are easily identifiable by their aquatic to terrestrial boundaries, but there are wide ranges of features that influence the nature of a particular lake (Figure 1-4). Several of these features will be described in this introduction.

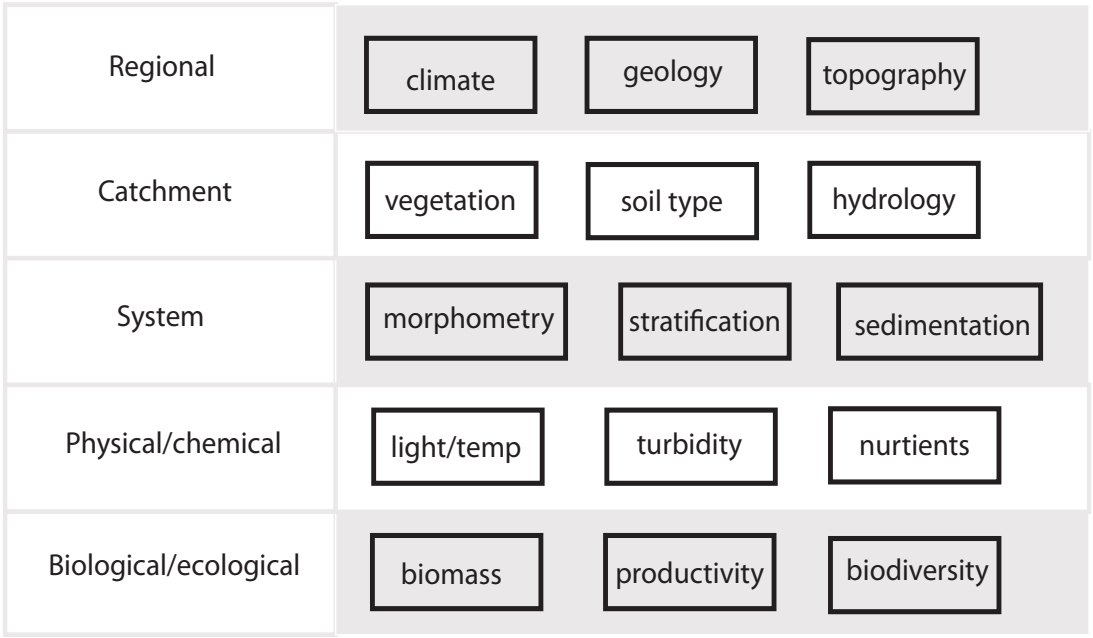

Figure 1-4: Hierarchical attributes affecting lake ecosystems. Adapted from Kalff, 2002

Hydrology and Landscape

As precipitation falls over terrestrial landscapes, it can take three routes into a lake (Figure 1-5). First, it can fall directly into the lake, importing nutrients and immigrant microorganisms with it. Precipitation can also seep into the groundwater table and eventually reach a lake after traveling through an aquifer. Finally, surface runoff or draining from streams may feed directly into lakes.

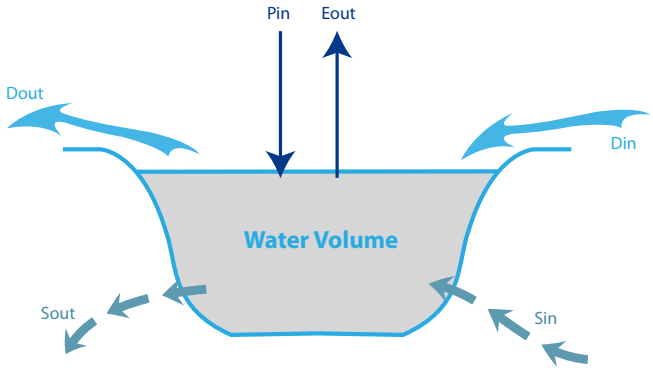

Figure 1-5: Water budget for lakes. Definitions: Pin, Precipitation in; Eout, Evaporation out; Din, Drainage in; Dout, Drainage out; Sin, Seepage in; Sout, Seepage out

Often, lakes are characterized as primarily seepage or drainage lakes. For example, Trout Bog Lake is a seepage lake, receiving no drainage or runoff, and Lake Mendota is a drainage lake, with several flowing inputs and outputs. Within drainage lakes, seepage can contribute up to 50% of the water inputs. Lakes that receive only inputs from precipitation also exist.

The primary input of water can have large ecological consequences (Kratz et al. 1997). Rivers and streams have the potential to transport large quantities of terrestrially derived nutrients and organic matter into lakes. The landscape position of lake within a catchment also potentially influences nutrient and carbon concentrations. Lakes that are higher in the landscape receive a higher proportion of precipitation which has an effect on cation and organic content (Kratz et al. 1997). If the lake's catchment is located in an agriculture region, there is a much larger contribution of nutrient originating from fertilizer runoff. Surface runoff also increases the turbidity of lakes, which decreases the penetration of photosynthetically active radiation.

Water chemistry is also significantly affected by retention time, which is determined by discharge rates and volume. Both primary productivity and respiration are inversely correlated to retention time.

Further, the lasting affect of bacterial immigrants appears to be controlled by retention time, where lakes with shorter retention times were composed of more terrestrial derived organisms (Lindstrom et al. 2006).

Temperature and Stratification

Limnologists have long recognized the effect of temperature in structuring biological communities. Strong thermal density stratification can create multiple uniform and mixed environments within a lake. In regions where climatic influences create strong seasonal temperature cycles, like the temperate North America, lakes typically become stratified into three layers during the summer months of high atmospheric heat and solar radiation coupled with wind induced currents (Figure 1-6).

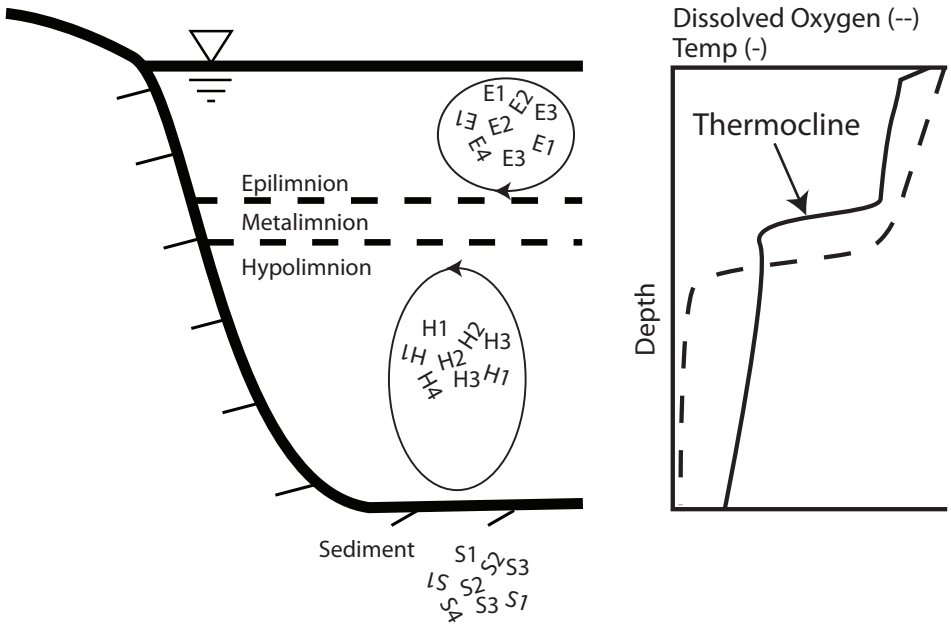

Figure 1-6: A lake profile during stratification. Each layer has a distinct microbial community and is considered well mixed

The hypolimnion is the bottom most layer during stratification. The hypolimnion is characterized by low temperatures (approximately 4 degrees Celsius), depleted oxygen, and laminar movement. Since

water is densest at 4 degrees Celsius, this bottom layer remains uniformly mixed, but no significant mixing occurs across the overlying thermal layers. The hypolimnion receives sinking detritus from production occurring in the overlying layers, and thus electron-accepting molecules are rapidly used during the decomposition of this organic matter. Finally, because there is a barrier between the hypolimnion and the atmosphere, there are no strong lateral forces (i.e. wind) that are causing turbulence within this environment, and thus the hypolimnion remains relatively still.

Directly above the hypolimnion lies the metalimnion. This layer spans form the top of the hypolimnion to the bottom of the epilimnion, over the depths at which rapid temperature decline is occurring. The location of rapid temperature decline is referred to as the thermocline. This layer causes the restriction of nutrient and gas fluxes across the epilimnion and hypolimnion, and it also hinders the ability of buoyant organisms to efficiently move between layers.

The epilimnion is the upper most mixed layer of a stratified lake. Here, photosynthetically active radiation is easily captured, and the warmer temperatures stimulate primary production in this layer. Primary production in this layer is usually dominates respiration, and it provides an energy source for the rest of the food web. This layer is more rapidly mixed than the hypolimnion due to wind forces, and is typically oxygenated. Additionally, this layer collects precipitation and any deposits of nutrients that come with it rainfall. However, rapid uptake of nutrients during primary production depletes the dissolved nutrient pool in this layer.

Lakes become stratified when the resistance to mixing water of different densities becomes larger than the mixing power from wind induced turbulence (Kalff 2002). This process is reversed, however, when seasons change and surface temperatures cool so that the density difference between layers is no longer large enough to resist wind induced mixing. Lakes can have varying mixing regimes, depending mostly on the regional climate and lake depth (Lewis 1983). In temperate climates, lakes with adequate depth are most often dimictic, meaning they mix during spring warming and fall cooling. If lakes are

shallow enough, the lakes can be polymictic, meaning that they mix on more than two occasions throughout the ice-free season. Meromictic lakes, alternatively, never completely mix due to chemicline that creates an upper mixolimnion and a lower monimolimnion. There is epilimnion, and hypolimnion stratification within the mixolimnion, but mixing across the chemicline is slow enough to maintain the chemical gradient (Boehrer and Schultze 2008).

Carbon Cycle

Carbon can enter a lake in two ways. The first is via primary production (autochthonous C). As atmospheric carbon is fixed into phytoplankton biomass (or into rooted plants), a new DOC pool becomes available for the food web. Zooplankton and other grazers take prey upon phytoplankton blooms, and members higher on the food web subsequently prey upon these grazers. Grazers do not consume all phytoplankton biomass, and some biomass ends up as detritus. Heterotrophic bacteria decompose this sinking organic matter, acquiring the carbon and nutrients it contains, and recycles it back into the food web. Bacterial decomposition will consume oxygen if it is available, and oxygen becomes depleted (creating an oxycline) at a depth that is comparable to the thermocline. Below the oxycline, the decomposition slows due to less favorable thermodynamics.

Autochthonous production can be supplemented by the addition of carbon that was fixed in terrestrial ecosystems (allochthonous C). Often, the allochthonous carbon is in a labile form that is an important pool of respireable carbon (del Giorgio and Peters 1994). This linkage of terrestrial and aquatic carbon cycles can help explain the net respiration of an aquatic ecosystem if it is relatively unproductive (Cole et al. 1994). Studies have shown that the overall mass accumulation and storage of autochthonous carbon in aquatic ecosystems is comparable what is observed in soils, but this is occurring with lower

rates of production. Thus, these works illustrate the relatively inefficient nature of aquatic microbes at respiring, particularly in the anaerobic hypolimnion.

Nutrients and Trophic Status

The trophic status of lakes is used as index to gauge nutrient concentrations, clarity, and productivity levels of a lake (Table 1-2). Oligotrophic lakes are characterized as nutrient poor, mesotrophic are characterized as having moderate nutrient levels, and eutrophic are characterized as being nutrient rich. The term dystrophic was subsequently coined to describe lakes with high humic content. These lakes have high overall nutrient levels but are poorly drained and there is typically less productivity and respiration. Considering all trophic levels, phosphorus and nitrogen are typically the most limiting nutrients within a lake, though phosphorus more often limits primary producers (source Schindler 1978). Thus, these nutrients are generally considered the main identifiers of trophic status within a lake.

Table 1-1: Trophic Status Characteristics. Values are given in micrograms per liter

	Total N	Total P	Chl-a	Clarity (m)
Oligotrophic	< 350	< 10	< 3.5	> 12
Mesotrophic	350-650	10-30	3.5-9	> 6
Eutrophic	650-1200	30-100	9-25	1.5-3

The nutrient concentrations observed in lakes are factors of intrinsic and external loading. Intrinsic loads are a result of heterotrophic bacteria recycling the pool of nutrients back into the food web. External loads are allochthonous sources of nutrients that are inputted into the ecosystem. Both of these loads have great impact on the short term and long term metabolism and trophic status of lakes.

The epilimnion concentration of phosphorus (usually in the form of phosphates) is strongly driven by rapid phytoplankton uptake. Heterotrophic bacteria decompose dead phytoplankton and control the turnover of phosphorus, along with the aid of grazers who do not consume the entirety of their prey (Carpenter et al. 1987). A fraction of the decaying phytoplankton sinks to the hypolimnion and sediments during decomposition, and this export depletes the phosphorus concentrations in the epilimnion during the stratification period by transporting it to other layers (Levine, Stainton, and Schindler 1986). Ferrous minerals may trap phosphorus (if oxygen is present) at the sediment hypolimnion interface when decomposition occurs in the sediments, thus preventing recycling of this nutrient back to the epilimnion. If the sediment is anoxic, however, phosphorus has the chance to diffuse back to the epilimnion to be taken up by phytoplankton. During spring and fall turnover, however, the organic nutrients that are in the sediments are mixed by turbulence and released back into the water column. This process is repeated annually, and elevated nutrients in the spring are followed by rapid depletion before increased concentrations in the fall.

Nitrogen, which is found in organic and inorganic forms, is less of a limiting factor on primary production. For those lakes that are limited by nitrogen, primary production is dominated by cyanobacteria. This bacterial phylum is able to supplement its nitrogen requirements by fixing atmospheric nitrogen into organic forms. Any subsequent nitrification will produce nitrate, which can sink to anoxic water before being denitrified into nitrogen gas. Thus, while phosphorus is not imported or exported out of the lake via the atmosphere, nitrogen may be gained or lost via atmospheric exchange. Nitrogen follows the same seasonal trends as phosphorus, due to similar processes that mix up sediments and release stored organic nitrogen during the spring and fall mixing events.

Extrinsic nutrient loading is primarily from drainage of the watershed, although direct precipitation and seepage can introduce allochthonous nutrients. During ice melts and precipitation events, there is increased erosion in the watershed and sediments are deposited into lakes. Ice and snow

melting events, in particular, play an important role in supplementing spring nutrient concentration that can stimulate the growth of phytoplankton.

Aquatic Microbial Ecology

Aquatic microbial ecology can be split into two primary areas of focus. The first is based upon bacterial communities, the second based on populations. A community in this sense is defined as the entirety of all populations inhabiting a delineated ecosystem. A population is defined as the entirety of an individual species inhabiting a delineated ecosystem.

There are considerations that must be made with these definitions. First, the definition of a species has historically been at debate, particularly at times with rapid advancement of tools that can be applied to evolutionary studies (Konstantinidis, et al. 2006; Stackebrandt and Goebel 1994). The definition of a true species can, perhaps, best be defined as a unit of distinct ecological relevance. Note that while this definition implies a functional coherence across a species, current methods of assessing species diversity with clusters of closely related 16S sequences as surrogates for species do not guarantee functional homogeneity across clusters. Nonetheless, the adoption of a 97% similarity across the 16S gene has been adopted as a standard indicator of a species.

The delineation of an ecosystem can also be problematic, but it is usually defined on a case-specific basis. For example, the atmosphere and sediment act as natural boundaries of lake ecosystems, thus a bacterial community could be defined as existing within these limits. Upon assessing this delineation, one could see that the bacterial communities would be heterogeneous with depth. The heterogeneity is greatest across thermal layers of a lake, and coherent communities can generally be observed within the epilimnion and hypolimnion of lakes (Ochs et al. 1995). Thus, for the purposes of aquatic ecology, bacterial communities are often defined as being epilimnetic or hypolimnetic (Figure 1-6).

Community Ecology

Microbial community ecology is the study of the structure of bacterial communities, as they inhabit a particular environment. This subject often focuses on the patterns of structure with time and space, with particular interest given to how those patterns correlate to environmental characteristics. The concepts of community stability and response to disturbance are often considered as well.

Compared to bacterial communities, the dynamics of phytoplankton have been well studied historically and are well understood (Sommer et al. 1986). Capturing the dynamics of bacterial communities has been a more recent endeavor, due to the availability of community fingerprinting techniques and high-throughput sequencing. Consistent with the fact that aquatic systems have fairly predictable physical and resource availability patterns on an annual scale based on observations from the North Temperate Lakes Long Term Ecological Research program (lter.limnology.wisc.edu), early studies found distinct seasonal bacterial succession in single year studies (Höfle et al. 1999) and multiyear studies (Kent et al. 2004). More robust studies spanning over ten years have identified repeated seasonal patterns that identify spring as more complex than summer and fall in terms of species co-occurrence (Kara et al. 2013).

Within a small humic lake, the species richness (the number of species identified in a sample) was observed to drop once in early summer corresponding to an increase in mixotrophic and heterotrophic flagellates, and again in the late summer corresponding to specific dinoflagellate blooms (Kent et al. 2004). Indeed, bacterial-phytoplankton interactions have been identified as considerable contributors to BCC in subsequent studies (Kent et al. 2007), yet this same study identified environmental characteristics as stronger drivers of composition.

Quality of organic carbon, as it relates to allochthonous and autochthonous sources, is one environmental factor that can drive community composition (Jones et al. 2009). Using chlorophyll *a* and watercolor measurements as indicators of the relative contribution of allochthonous Jones et al. correlated

organic carbon quality to community composition in a multi-lake study. Crump et al. also observed community shifts upon the input of allochthonous carbon into an arctic lake during seasonal changes (Crump et al. 2003). Mesocosm experiments observing the community shifts upon additions of DOC of varying quality have supported the findings made in spatial and temporal studies (Eiler et al. 2003).

In addition to the quality of carbon, many other environmental factors influence BCC. Temperature and dissolved oxygen are two significant drivers of habitat quality and the community structure (Shade et al. 2008). Nutrient addition experiments have also demonstrated the affect of N, P, and C on BCC, though these additions had seasonally dependent responses (Newton and McMahon 2011).

Spatial and geographical patterns of bacterial composition have also been identified. Regionalism was identified in lakes spanning across Wisconsin in regards to certain OTUs and community richness, and drainage-versus-seepage classification had statistical correlations to BCC (Yannarell and Triplett 2005). The regionalism identified was presumably due to variances among local vegetation and climate patterns. Hydrology has also been documented as a physical factor that has correlation to specific community structures, as seen in lakes with differing retention times (Lindstrom and Bergstrom 2004).

Ultimately, the drivers of BCC are immense. While no models have been able to successfully predict BCC given environmental, biological, and physical parameters, the BCC of lakes tend to repeat seasonal and annual patterns. Repeatable responses to disturbances in aquatic ecosystems are also seen. Bacterial communities respond to mixing events differentially, depending upon the strength of mixing, yet aquatic communities show resilience upon these disturbances (Shade et al. 2010).

Advanced sequencing technology is revolutionizing the field of community ecology. Huge strides in temporal and spatial patterns can be made because of the affordability of Illumina and pyrosequencing (Gonzalez et al. 2012). The true diversity of the rare biosphere can now be more adequately assessed, which is changing the way we interpret community composition.

Population Ecology

The goal of population level ecology is to identify the drivers, patterns, and ecological function of a bacterial population. Inferences can be made from abundance data, genomic data, among other investigative data. Population ecology is becoming much more active given the depth of sequencing data that is being produced, but the bulk of knowledge has been built around the abundant and ubiquitous aquatic bacteria via other culture-independent techniques.

Before high-throughput sequencing, it was difficult to assign taxonomy to 'fingerprints' using methods such as ARISA and T-RFLP and laborious to sequence enough clonal 16S genes to track rare OTUs. For these methodological reasons, and for the simple fact that numerical dominance implies ecological relevance, early population level studies were performed on the abundant and ubiquitous taxa. Such taxa include the acI lineage (roughly equivalent to genus, see (Newton et al. 2011) for explanation on phylogenetic definitions), *Polynucleobacter*, and the LD12 tribe.

Belonging to the Actinobacteria phyla, acI is a persistently abundant freshwater bacteria with few dynamics across seasons (Glöckner et al. 2000). No isolates from this clade exist, but single cell amplified genomes of the acI-B1 tribe showcased a very small genome of approximately 1.2 Mbp (Garcia et al. 2013). The apparent success of this tribe may stem from the ability to harness light energy via rhodopsin genes, though the rest of the genome suggests a facultative aerobic strategy (Garcia et al. 2013; Sharma et al. 2008).

Almost equally as ubiquitous, the *Polynucleobacter* genus (Betaproteobacteria) is perhaps the most well studied freshwater taxonomic group to date (Newton et al. 2011). Cultured representatives have been obtained with relative success. Species have been successfully grown on agar plates with the addition of organic acids (Watanabe et al. 2009), but more laborious effort has gone into effective isolation of many strains based on substrate preferences (Hahn et al. 2012; Hahn et al. 2009). While

genomes and substrate preferences have been studied, species within this genus seem to have distribution patterns that are particularly dependent upon pH (Wu et al. 2006). Additionally, specific species appear to be enriched by allochthonous sources of carbon (Hutalle-Schmelzer et al. 2010).

LD12 (Alphaproteobacteria) is perhaps the most widely distributed freshwater taxon. This tribe demonstrates strong preference towards glutamine and glutamate, and its success seems to be correlated greatly with seasonal temperature increases (Salcher et al. 2011). As was identified for the acI-B1, single cell amplified genomes of LD12 contained rhodopsin like genes (Martinez-Garcia et al. 2011), which may be one explanation for its dominance in oligotrophic lakes.

Each of these taxa has long been identified as numerically dominant, but specific ecological roles have been difficult to interpret. The inclusion of genomic data from single amplified genomes is altering the way we interpret the perceived success of bacterial species and how they interact with their environment. Within just the last few years, knowledge on the diverse metabolic strategies that each population encodes has given researches better insights into their ecological success. The application of 16S tag sequencing has made it possible to acquire long-term temporal abundance patterns for abundant and rare species.

In one such study, aquatic populations were investigated for co-occurring patterns. Eiler et al. took several different approaches to cluster OTUs with shared dynamics, and subsequently identified the environmental factors that correlated with the peak abundances of various clusters (Eiler et al. 2011). Using k-means clustering, certain groups of taxa were found to be at peak abundance during zooplankton blooms, cyanobacteria blooms, and diatom blooms. This work was especially enlightening as high-throughput sequencing was used, and the 50 most abundant OTUs could be considered in the analysis.

With the onset of sequencing technology and advanced cell sorting capabilities, researchers are rapidly gaining ground on the “who’s who” of aquatic bacteria. Unpublished genome sequences from single cell sorting projects are currently being analyzed for functional characteristics. In particular,

freshwater microbial ecologists are taking advantage of these techniques to identify what allows the abundant taxa to be numerically dominant. Deep 16S tag sequencing is also enhancing researchers' ability to identify the core and rare communities that are ubiquitous across time and space. In studies involving time-series of data, researchers are able to correlate environmental characteristics with species abundance patterns, with the ultimate goal of predicting how bacterial communities will behave in all environmental conditions.

References

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. 2013. "Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes." *Nature Biotechnology* 31 (6) (June): 533–8.

Altenhoff, Adrian M, and Christophe Dessimoz. 2009. "Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods." *PLoS Computational Biology* 5 (1) (January): e1000262.

Altschup, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215: 403–410.

Amann, R I, W Ludwig, and K H Schleifer. 1995. "Phylogenetic Identification and in Situ Detection of Individual Microbial Cells Without Cultivation." *Microbiological Reviews* 59 (1) (March): 143–69.

Aparicio, Samuel, Jarrod Chapman, Elia Stupka, Nik Putnam, Jer-Ming Chia, Paramvir Dehal, Alan Christoffels, et al. 2002. "Whole-genome Shotgun Assembly and Analysis of the Genome of Fugu Rubripes." *Science* 297 (5585) (August 23): 1301–10.

Azad, Rajeev K, and Mark Borodovsky. 2004. "Probabilistic Methods of Identifying Genes in Prokaryotic Genomes: Connections to the HMM Theory." *Briefings in Bioinformatics* 5 (2) (June): 118–30.

Azam, F., T. Fenchel, J. G. Field, J. S . Gray, L. A. Meyer-Reil, and F. Thingstad. 1983. "The Ecological Role of Water-Column Microbes in the Sea." *Marine Ecology* 10: 257–263.

Badger, J H, and G J Olsen. 1999. "CRITICA: Coding Region Identification Tool Invoking Comparative Analysis." *Molecular Biology and Evolution* 16 (4) (April): 512–24.

Bairoch, a. 2000. "The ENZYME Database in 2000." *Nucleic Acids Research* 28 (1) (January 1): 304–5.

Barea, José-Miguel, María José Pozo, Rosario Azcón, and Concepción Azcón-Aguilar. 2005. "Microbial Co-operation in the Rhizosphere." *Journal of Experimental Botany* 56 (417) (July): 1761–78.

Batzogulou, Serafim, David B Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P Mesirov, and Eric S Lander. 2002. "ARACHNE: A Whole-Genome Shotgun Assembler." *Genome Research* 12 (July): 177–189. doi:10.1101/gr.208902.

Behrens, Sebastian, Tina Lösekann, Jennifer Pett-Ridge, Peter K Weber, Wing-On Ng, Bradley S Stevenson, Ian D Hutcheon, David a Relman, and Alfred M Spormann. 2008. "Linking Microbial Phylogeny to Metabolic Activity at the Single-cell Level by Using Enhanced Element Labeling-catalyzed Reporter Deposition Fluorescence in Situ Hybridization (EL-FISH) and NanoSIMS." *Applied and Environmental Microbiology* 74 (10) (May): 3143–50.

Berg, R D. 1996. "The Indigenous Gastrointestinal Microflora." *Trends in Microbiology* 4 (11) (November): 430–5.

Besemer, John, and Mark Borodovsky. 2005. “GeneMark: Web Software for Gene Finding in Prokaryotes, Eukaryotes and Viruses.” *Nucleic Acids Research* 33 (Web Server issue) (July 1): W451–4. doi:10.1093/nar/gki487.

Boehrer, Bertram, and Martin Schultze. 2008. “Stratification of Lakes.” *Reviews of Geophysics* 46 (2006):

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High- Throughput Community Sequencing Data I.” *Nature Publishing Group* 7 (5): 335–336.

Carpenter, S R, J F Kitchell, J R Hodgson, P A Cochran, J J Elser, M M Elser, D M Lodge, D Kretchmer, X He, and C N von Ende. 1987. “Regulation of Lake Primary Productivity by Food Web Structure.” *Ecology* 68 (6): 1863–1876.

Chen, I-Min A., Victor M Markowitz, Ken Chu, Iain Anderson, Konstantinos Mavromatis, Nikos C Kyrpides, and Natalia N Ivanova. 2013. “Improving Microbial Genome Annotations in an Integrated Database Context.” *PloS One* 8 (2) (January): e54859.

Cole, J J, N F Caraco, G W Kling, and T K Kratz. 1994. “Carbon Dioxide Supersaturation in the Surface Waters of Lakes.” *Science* 265 (5178) (September 9): 1568–70.

Cole, J R, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M Mcgarrell, T Marsh, and G M Garrity. 2009. “The Ribosomal Database Project: Improved Alignments and New Tools for rRNA Analysis.” *Nucleic Acids Research* 37 (Database issue) (January): D141–5.

Cotner, James B, and Bopaiah A Biddanda. 2002. “Small Players, Large Role: Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems.” *Ecosystems* 5 (2): 105–121.

Crump, Byron C, George W Kling, Michele Bahr, and John E Hobbie. 2003. “Bacterioplankton Community Shifts in an Arctic Lake Correlate with Seasonal Changes in Organic Matter Source.” *Applied and Environmental Microbiology* 69 (4): 2253–2268.

DeAngelis, Kristen M., John M. Gladden, Martin Allgaier, Patrik D’haeseleer, Julian L. Fortney, Amitha Reddy, Philip Hugenholtz, et al. 2010. “Strategies for Enhancing the Effectiveness of Metagenomic-based Enzyme Discovery in Lignocellulolytic Microbial Communities.” *BioEnergy Research* 3 (2) (March 30): 146–158.

Degnan, Patrick H, and Howard Ochman. 2011. “Illumina-based Analysis of Microbial Community Diversity.” *The ISME Journal* 6 (1) (June 16): 183–194.

Delong, Edward F, Gene S Wickham, and Norman R Pace. 1989. “Phylogenetic Stains : Ribosomal RNA-Based Probes for the Identification of Single Cells” 243 (4896): 1360–1363.

DeSantis, T Z, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. 2006. “Greengenes, a Chimera-checked 16S rRNA Gene Database and Workbench Compatible with ARB.” *Applied and Environmental Microbiology* 72 (7) (July): 5069–72.

DeSantis, Todd Z, Keith Keller, Ulas Karaoz, Alexander V Alekseyenko, Navjeet N S Singh, Eoin L Brodie, Zhiheng Pei, Gary L Andersen, and Niels Larsen. 2011. "Simrank: Rapid and Sensitive General-purpose K-mer Search Tool." *BMC Ecology* 11 (1) (January): 11.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than BLAST." *Bioinformatics* 26 (19) (October 1): 2460–2461.

Edgar, Robert C, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. 2011. "UCHIME Improves Sensitivity and Speed of Chimera Detection." *Bioinformatics (Oxford, England)* 27 (16) (August 15): 2194–200.

Eiler, Alexander, and Stefan Bertilsson. 2007. "Flavobacteria Blooms in Four Eutrophic Lakes: Linking Population Dynamics of Freshwater Bacterioplankton to Resource Availability." *Applied and Environmental Microbiology* 73 (11) (June): 3511–8.

Eiler, Alexander, Friederike Heinrich, and Stefan Bertilsson. 2011. "Coherent Dynamics and Association Networks Among Lake Bacterioplankton Taxa." *The ISME Journal* (September 1): 1–13.

Eiler, Alexander, Silke Langenheder, Stefan Bertilsson, and Lars J Tranvik. 2003. "Heterotrophic Bacterial Growth Efficiency and Community Structure at Different Natural Organic Carbon Concentrations." *Applied and Environmental Microbiology* 69 (7): 3701–3709.

Eren, A. Murat, Marcela Zozaya, Christopher M. Taylor, Scot E. Dowd, David H. Martin, and Michael J. Ferris. 2011. "Exploring the Diversity of Gardnerella Vaginalis in the Genitourinary Tract Microbiota of Monogamous Couples Through Subtle Nucleotide Variation." *PloS One* 6 (10) (January): e26732.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. "Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment." *Genome Research* 8 (3) (March 1): 175–185.

Felsenstein, Joseph, and Gary A Churchill. 1996. "A Hidden Markov Model Approach Evolution to Variation Among Sites in Rate of Evolution." *Molecular Biology and Evolution* 13 (1): 93–104.

Fierer, Noah, Christy M McCain, Patrick Meir, Michael Zimmermann, Joshua M Rapp, Miles R Silman, and Rob Knight. 2011. "Microbes Do Not Follow the Elevational Diversity Patterns of Plants and Animals." *Ecology* 92 (4): 797–804.

Fischer, S G, and L S Lerman. 1983. "DNA Fragments Differing by Single Base-pair Substitutions Are Separated in Denaturin Ggradient Gels: Correspondence with Melting Theory." *Proceedings of the National Academy of Sciences* 80 (March): 1579–1583.

Fisher, Madeline M, and Eric W Triplett. 1999. "Automated Approach for Ribosomal Intergenic Spacer Analysis of Microbial Diversity and Its Application to Freshwater Bacterial Communities." *Applied and Environmental Microbiology* 65 (10): 4630–4636.

Fleischmann, R D, M D Adams, O White, R a Clayton, E F Kirkness, a R Kerlavage, C J Bult, J F Tomb, B a Dougherty, and J M Merrick. 1995. "Whole-genome Random Sequencing and Assembly of Haemophilus Influenzae Rd." *Science* 269 (5223) (July 28): 496–512.

Francisco, Donald E, Robert A Mah, and Albert C Rabin. 1973. “Acridine Orange-Epifluorescence Technique For Couting Bacteria in Natural Waters” 92 (3): 416–421.

Frishman, D, a Mironov, H W Mewes, and M Gelfand. 1998. “Combining Diverse Evidence for Gene Recognition in Completely Sequenced Bacterial Genomes.” *Nucleic Acids Research* 26 (12) (June 15): 2941–7.

Fuhrman, J A, J W Ammerman, and F Azam. 1980. “Bacterioplankton in the Coastal Euphotic Zone: Distribution, Activity and Possible Relationships with Phytoplankton.” *Marine Biology* 60 (2-3): 201–207.

Fuhrman, J A, and F Azam. 1982. “Thymidine Incorporation as a Measure of Heterotrophic Bacterioplankton Production in Marine Surface Waters: Evaluation and Field Results.” *Marine Biology* 66 (2): 109–120.

Galand, Pierre E, Emilio O Casamayor, David L Kirchman, Marianne Potvin, and Connie Lovejoy. 2009. “Unique Archaeal Assemblages in the Arctic Ocean Unveiled by Massively Parallel Tag Sequencing.” *The ISME Journal* 3 (7) (July): 860–9.

Garcia, Sarahi L, Katherine D McMahon, Manuel Martinez-Garcia, Abhishek Srivastava, Alexander Sczyrba, Ramunas Stepanauskas, Hans-Peter Grossart, Tanja Woyke, and Falk Warnecke. 2013. “Metabolic Potential of a Single Cell Belonging to One of the Most Abundant Lineages in Freshwater Bacterioplankton.” *The ISME Journal* 7 (1) (January): 137–47.

Gerlach, Wolfgang, Sebastian Jünemann, Felix Tille, Alexander Goesmann, and Jens Stoye. 2009. “WebCARMA: a Web Application for the Functional and Taxonomic Classification of Unassembled Metagenomic Reads.” *BMC Bioinformatics* 10 (January): 430. doi:10.1186/1471-2105-10-430.

Gilbert, Jack A, Folker Meyer, Janet Jansson, Jeff Gordon, Norman Pace, James Tiedje, Ruth Ley, et al. 2010. “The Earth Microbiome Project : Meeting Report of the ‘ 1 St EMP Meeting on Sample Selection and Acquisition ’ at Argonne National Laboratory October 6 Th 2010 .” *Standards in Genomic Sciences* 3 (3): 249–253.

Del Giorgio, Paul A, and Robert H Peters. 1994. “In Lakes : Influence in Planktonic Patterns P : R Ratios of Lake Trophy and Dissolved Organic Carbon.” *Limnology and Oceanography* 39 (4): 772–787.

Glenn, Travis C. 2011. “Field Guide to Next-generation DNA Sequencers.” *Molecular Ecology Resources* 11 (5) (September): 759–69.

Glöckner, Frank Oliver, Evgeny Zaichikov, Natalia Belkova, Ludmilla Denissova, Jakob Pernthaler, Annelie Pernthaler, and Rudolf Amann. 2000. “Comparative 16S rRNA Analysis of Lake Bacterioplankton Reveals Globally Distributed Phylogenetic Clusters Including an Abundant Group of Actinobacteria.” *Applied and Environmental Microbiology* 66 (11): 5053–5065.

Gonzalez, Antonio, Andrew King, Michael S Robeson, Sejin Song, Ashley Shade, Jessica L Metcalf, and Rob Knight. 2012. “Characterizing Microbial Communities Through Space and Time.” *Current Opinion in Biotechnology* 23 (3) (June): 431–6.

Haas, Brian J, Dirk Gevers, Ashlee M Earl, Mike Feldgarden, Doyle V Ward, Georgia Giannoukos, Dawn Ciulla, et al. 2011. "Chimeric 16S rRNA Sequence Formation and Detection in Sanger and 454-pyrosequenced PCR Amplicons." *Genome Research* 21 (3) (March): 494–504.

Hahn, Martin W, Elke Lang, Ulrike Brandt, Qinglong L Wu, and Thomas Scheuerl. 2009. "Emended Description of the Genus Polynucleobacter and the Species Polynucleobacter Necessarius and Proposal of Two Subspecies , P . Necessarius Subsp . Necessarius Subsp . Nov . And." *International Journal of Systematic and Evolutionary Microbiology*: 2002–2009.

Hahn, Martin W, Arevik Minasyan, Elke Lang, and Ulrike Koll. 2012. "Polynucleobacter Difficilis Sp . Nov ., a Planktonic Freshwater Bacterium Affiliated with Subcluster B1 of the Genus Polynucleobacter." *International Journal of Systematic and Evolutionary Microbiology*: 376–383.

Handelsman, Jo, Michelle R Rondon, Sean F Brady, Jon Clardy, and Robert M Goodman. 1998. "Molecular Biological Access to the Chemistry of Unknown Soil Microbes: a New Frontier for Natural Products." *Chemistry and Biology* 5 (10): R245–R249.

Hao, Xiaolin, and Ting Chen. 2012. "OTU Analysis Using Metagenomic Shotgun Sequencing Data." *PloS One* 7 (11) (January): e49785.

Heath, Lenwood S, and Amrita Pati. 2007. "Genomic Signatures in De Bruijn Chains": 216–227.

Holt, John G, and Peter H Sneath. 2005. *Bergey's Manual of Systematic Bacteriology*. Ed. Don J Brenner, Noel R Krieg, and James T Staley. Second. Vol. 2. Springer.

Huse, Susan M, Les Dethlefsen, Julie a Huber, David Mark Welch, David Mark Welch, David a Relman, and Mitchell L Sogin. 2008. "Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing." *PLoS Genetics* 4 (11) (November): e1000255.

Huse, Susan M, David Mark Welch, Hilary G Morrison, and Mitchell L Sogin. 2010. "Ironing Out the Wrinkles in the Rare Biosphere Through Improved OTU Clustering." *Environmental Microbiology* 12 (7) (July): 1889–98.

Huson, Daniel H, Alexander F Auch, Ji Qi, and Stephan C Schuster. 2007. "MEGAN Analysis of Metagenomic Data." *Genome Research* 17 (3) (March): 377–86.

Hutalle-Schmelzer, Kristine Michelle L, Elke Zwirnmann, Angela Krüger, and Hans-Peter Grossart. 2010. "Enrichment and Cultivation of Pelagic Bacteria from a Humic Lake Using Phenol and Humic Matter Additions." *FEMS Microbiology Ecology* 72 (1) (April): 58–73.

Höfle, Manfred G, Heike Haas, and Katja Dominik. 1999. "Seasonal Dynamics of Bacterioplankton Community Structure in a Eutrophic Lake as Determined by 5S rRNA Analysis." *Applied and Environmental Microbiology* 65 (7).

Jones, Stuart E, Ryan J Newton, and Katherine D McMahon. 2009. "Evidence for Structuring of Bacterial Community Composition by Organic Carbon Source in Temperate Lakes." *Environmental Microbiology* 11 (9) (September): 2463–72.

Kalff, J. 2002. *Limnology: Inland Water Ecosystems*. Prentice Hall.

Kanehisa, M, and S Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1) (January 1): 27–30.

Kara, Emily L, Paul C Hanson, Yu Hen Hu, Luke Winslow, and Katherine D McMahon. 2013. “A Decade of Seasonal Dynamics and Co-occurrences Within Freshwater Bacterioplankton Communities from Eutrophic Lake Mendota, WI, USA.” *ISME J* 7 (3) (March): 680–684.

Kent, A D, S E Jones, A C Yannarell, J M Graham, G H Lauster, T K Kratz, and E W Triplett. 2004. “Annual Patterns in Bacterioplankton Community Variability in a Humic Lake.” *Microbial Ecology* 48 (4) (November): 550–60.

Kent, A D, A C Yannarell, J A Rusak, E W Triplett, and Katherine D McMahon. 2007. “Synchrony in Aquatic Microbial Community Dynamics.” *The ISME Journal* 1 (1) (May): 38–47.

Kerstens, H. M., P. J. Poddighe, and a. G. Hanselaar. 1995. “A Novel in Situ Hybridization Signal Amplification Method Based on the Deposition of Biotinylated Tyramine.” *Journal of Histochemistry & Cytochemistry* 43 (4) (April 1): 347–352.

Konstantinidis, Konstantinos T, Alban Ramette, and James M Tiedje. 2006. “The Bacterial Species Definition in the Genomic Era.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361 (1475) (November 29): 1929–40.

Kostka, Joel E, Om Prakash, Will A Overholt, Stefan J Green, Gina Freyer, Andy Canion, Jonathan Delgardio, Nikita Norton, Terry C Hazen, and Markus Huettel. 2011. “Hydrocarbon-Degrading Bacteria and the Bacterial Community Response in Gulf of Mexico Beach Sands Impacted by the Deepwater Horizon Oil Spill †‡.” *Society* 77 (22): 7962–7974.

Kratz, Timothy K., Katherine E. Webster, Carl J. Bowser, John J. Magnuson, and Barbara J. Benson. 1997. “The Influence of Landscape Position on Lakes in Northern Wisconsin.” *Freshwater Biology* 37: 209–217.

Krzmarzick, Mark J, Patrick J McNamara, Benjamin B Crary, and Paige J Novak. 2013. “Abundance and Diversity of Organohalide-respiring Bacteria in Lake Sediments Across a Geographical Sulfur Gradient.” *FEMS Microbiology Ecology* 84 (2) (May): 248–58.

Kunin, Victor, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. 2008. “A Bioinformatician’s Guide to Metagenomics.” *Microbiology and Molecular Biology Reviews : MMBR* 72 (4) (December): 557–78, Table of Contents.

Lapidus, Alla L. 2009. “Genome Sequence Databases (Overview): Sequencing and Assembly.” *Lawrence Berkely National Laboratory*.

Levine, S N, M P Stainton, and D W Schindler. 1986. “A Radiotracer Study of Phosphorus Cycling in a Eutrophic Canandian Shield Lake, Lake 227, Northwestern Ontario.” *Can. J. Fish. Aquat. Sci.* 43: 366–78.

Lewis, William M. Jr. 1983. “A Revised Classification of Lakes Based on Mixing.” *Can. J. Fish. Aquat. Sci.* 40: 1779–1787.

Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, et al. 2010. "De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing." *Genome Research* 20 (2) (February): 265–72.

Lindeman, R L. 1942. "The Trophic-Dynamic Aspect of Ecology." *Ecology* 23 (4): 399–417.

Lindstrom, Eva S, and A. K. Bergstrom. 2004. "Influence of Inlet Bacteria on Bacterioplankton Assemblage Composition in Lakes of Different Retention Time Hydraulic." *Limnology and Oceanography* 49 (1): 125–136.

Lindstrom, Eva S., Markus Forslund, Grete Algest, and Ann-Kristin Bergstrom. 2006. "External Control of Bacterial Community Structure in Lakes." *Limnology and Oceanography* 51 (1): 339–342.

Liu, Wen-tso, Terence L Marsh, Hans Cheng, and L J Forney. 1997. "Characterization of Microbial Diversity by Determining Terminal Restriction Fragment Length Polymorphisms of Genes Encoding 16S." *Applied and Environmental Microbiology* 63 (11): 4516–4522.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa a Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-density Picolitre Reactors." *Nature* 437 (7057) (September 15): 376–80.

Markowitz, Victor M, I-Min a Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Yuri Grechkin, Anna Ratner, et al. 2012. "IMG/M: The Integrated Metagenome Data Management and Comparative Analysis System." *Nucleic Acids Research* 40 (Database issue) (January): D123–9.

Martinez-Garcia, Manuel, Brandon K Swan, Nicole J Poulton, Monica Lluesma Gomez, Dashiell Masland, Michael E Sieracki, and Ramunas Stepanauskas. 2011. "High-throughput Single-cell Sequencing Identifies Photoheterotrophs and Chemoautotrophs in Freshwater Bacterioplankton." *The ISME Journal* (June 30): 1–11.

Mavromatis, Konstantinos, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eugene Goltsman, Alice C Mchardy, Isidore Rigoutsos, et al. 2007. "Use of Simulated Data Sets to Evaluate the Fidelity of Metagenomic Processing Methods." *Nature Methods* 4 (6): 495–500.

Mavromatis, Konstantinos, Natalia N Ivanova, I-Min A Chen, Ernest Szeto, M Victor, and Nikos C Kyrpides. 2009. "The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes." *Standards in Genomic Sciences* 1: 63–67.

Mazin, Pavel V, Mikhail S Gelfand, Andrey A Mironov, Aleksandra B Rakhmaninova, Anatoly R Rubinov, Robert B Russell, and Olga V Kalinina. 2010. "An Automated Stochastic Approach to the Identification of the Protein Specificity Determinants and Functional Subfamilies." *Algorithms for Molecular Biology: AMB* 5 (January): 29.

Meyer, F, D Paarmann, M D'Souza, R Olson, E M Glass, M Kubal, T Paczian, et al. 2008. "The Metagenomics RAST Server - a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes." *BMC Bioinformatics* 9 (January): 386.

Miller, Jason R, Sergey Koren, and Granger Sutton. 2010. "Assembly Algorithms for Next-generation Sequencing Data." *Genomics* 95 (6) (June): 315–27.

Mullis, Kary B., and Fred A. Faloona. 1987. “Specific Synthesis of DNA in Vitro via a Polymerase-Catalyzed Chain Reation.” *Methods in Enzymology* 155: 335–350.

Muyzer, G, E C de Waal, and a G Uitterlinden. 1993. “Profiling of Complex Microbial Populations by Denaturing Gradient Gel Electrophoresis Analysis of Polymerase Chain Reaction-amplified Genes Coding for 16S rRNA.” *Applied and Environmental Microbiology* 59 (3) (March): 695–700.

Newton, Ryan J, Stuart E Jones, Alexander Eiler, Katherine D McMahon, and Stefan Bertilsson. 2011. “A Guide to the Natural History of Freshwater Lake Bacteria.” *Microbiology and Molecular Biology Reviews : MMBR* 75 (1) (March): 14–49.

Newton, Ryan J, and Katherine D McMahon. 2011. “Seasonal Differences in Bacterial Community Composition Following Nutrient Additions in a Eutrophic Lake.” *Environmental Microbiology* 13 (4) (April): 887–99.

Oh, Seungdae, Alejandro Caro-Quintero, Despina Tsementzi, Natasha Deleon-Rodriguez, Chengwei Luo, Rachel Poretsky, and Konstantinos T Konstantinidis. 2011. “Metagenomic Insights into the Evolution, Function and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem.” *Applied and Environmental Microbiology* 77 (17) (July 15): 6000–6011.

Overbeek, Ross, Tadhg Begley, Ralph M Butler, Jomuna V Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, et al. 2005. “The Subsystems Approach to Genome Annotation and Its Use in the Project to Annotate 1000 Genomes.” *Nucleic Acids Research* 33 (17) (January): 5691–702.

Pace, N. R. 1997. “A Molecular View of Microbial Diversity and the Biosphere.” *Science* 276 (5313) (May 2): 734–740.

Pati, Amrita, Lenwood S Heath, Nikos C Kyrpides, and Natalia Ivanova. 2011. “ClaMS: A Classifier for Metagenomic Sequences.” *Standards in Genomic Sciences* 5 (2) (November 30): 248–53.

Patin, Nastassia V, Victor Kunin, Ulrika Lidström, and Matthew N Ashby. 2013. “Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates.” *Microbial Ecology* 65 (3) (April): 709–19.

Paul, John H. 1982. “Use of Hoechst Dyes 33258 and 33342 for Enumeration of Attached and Planktonic Bacteria.” *Applied and Environmental Microbiology* 43 (4): 939–944.

Paver, Sara F, and Angela D Kent. 2010. “Temporal Patterns in Glycolate-utilizing Bacterial Community Composition Correlate with Phytoplankton Population Dynamics in Humic Lakes.” *Microbial Ecology* 60 (2) (August): 406–18.

Pernthaler, Annelie, Jakob Pernthaler, and Rudolf Amann. 2002. “Fluorescence In Situ Hybridization and Catalyzed Reporter Deposition for the Identification of Marine Bacteria” 68 (6).

Pevzner, P a, H Tang, and M S Waterman. 2001. “An Eulerian Path Approach to DNA Fragment Assembly.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (17) (August 14): 9748–9753.

Pomeroy, Lawrence R. 1974. “The Ocean’s Food Web, A Changing Paradigm.” *BioSciences* 24 (9): 499–504.

Porter, K G, and Y S Feig. 1980. “The Use of DAPI for Identification and Enumeration of Bacteria and Blue-green Algae.” *Limnol. Oceanogr* 25: 943–948.

Pruesse, Elmar, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. 2007. “SILVA: a Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB.” *Nucleic Acids Research* 35 (21) (January): 7188–96.

Rappé, Michael S, and Stephen J Giovannoni. 2003. “The Uncultured Microbial Majority.” *Annual Review of Microbiology* 57 (January): 369–94.

Riley, Gordon Arthur. 1951. *Oxygen, Phosphate, and Nitrate in the Atlantic Ocean*. Bingham Oceanographic Laboratory.

Ronaghi, Mostafa, Mathias Uhlen, and Pal Nyren. 2013. “A Sequencing Real-Time Method Based on Pyrophosphate.” *Science* 281 (5375): 363–365.

Salcher, Michaela M, Jakob Pernthaler, and Thomas Posch. 2011. “Seasonal Bloom Dynamics and Ecophysiology of the Freshwater Sister Clade of SAR11 Bacteria ‘That Rule the Waves’ (LD12).” *The ISME Journal* 5 (8) (August): 1242–52.

Sanger, F, S Nicklen, and R Coulson. 1977. “DNA Sequencing with Chain-terminating.” *Proceedings of the National Academy of Sciences* 74 (12): 5463–5467.

Schbath, Sophie, Bernard Prum, and Elisabeth De Turckheim. 1995. “Exceptional Motifs in Different Markov Chain Models Statistical Analysis of DNA Sequences.” *Journal of Computational Biology* 2 (3): 417–437.

Schloss, Patrick D. 2010. “The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-based Studies.” *PLoS Computational Biology* 6 (7) (January): e1000844.

Schloss, Patrick D, and Jo Handelsman. 2005. “Introducing DOTUR , a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness.” *Applied and Environmental Microbiology* 71 (3): 1501–1506.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan a Lesniewski, et al. 2009. “Introducing Mothur: Open-source, Platform-independent, Community-supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23) (December): 7537–41.

Shade, A, S E Jones, and K D McMahon. 2008. “The Influence of Habitat Heterogeneity on Freshwater Bacterial Community Composition and Dynamics.” *Environmental Microbiology* 10 (4): 1057–1067.

Shade, Ashley, Gregory J Caporaso, Jo Handelsman, Rob Knight, and Noah Fierer. 2013. "A Meta-analysis of Changes in Bacterial and Archaeal Communities with Time." *The ISME Journal* (April 11): 1–14.

Shade, Ashley, Chih-Yu Chiu, and Katherine D McMahon. 2010. "Seasonal and Episodic Lake Mixing Stimulate Differential Planktonic Bacterial Dynamics." *Microbial Ecology* 59 (3) (April): 546–54.

Sharma, Adrian K, Olga Zhaxybayeva, R Thane Papke, and W Ford Doolittle. 2008. "Actinorhodopsins: Proteorhodopsin-like Gene Sequences Found Predominantly in Non-marine Environments." *Environmental Microbiology* 10 (4): 1039–1056.

Shih, Patrick M, Dongying Wu, Amel Latifi, Seth D Axen, David P Fewer, Emmanuel Talla, Alexandra Calteau, et al. 2013. "Improving the Coverage of the Cyanobacterial Phylum Using Diversity-driven Genome Sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 110 (3) (January 15): 1053–8.

Sommer, U, Z M Gliwicz, W Lampert, and A Duncan. 1986. "The {PEG-model} of Seasonal Succession of Planktonic Event in Fresh Waters." *Arch. Hydrobiol.* 106 (4): 433–471.

Sorek, Rotem, Yiwen Zhu, Christopher J Creevey, M Pilar Francino, Peer Bork, and Edward M Rubin. 2007. "Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer." *Science* 318 (5855): 1449–1452.

Stackebrandt, E., and B. M. Goebel. 1994. "Taxonomic Note: A Place for DNA-DNA Reassociation and 16s rRNA Sequence Analysis in the Present Species Definition in Bacteriology." *International Journal of Systematic Bacteriology* 44 (4) (October 1): 846–849.

Sun, Yijun, Yunpeng Cai, Susan M Huse, Rob Knight, William G Farmerie, Xiaoyu Wang, and Volker Mai. 2012. "A Large-scale Benchmark Study of Existing Algorithms for Taxonomy-independent Microbial Community Analysis." *Briefings in Bioinformatics* 13 (1) (January): 107–21.

Tatusov, R. L. 1997. "A Genomic Perspective on Protein Families." *Science* 278 (5338) (October 24): 631–637.

Teeling, Hanno, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. 2004. "Application of Tetranucleotide Frequencies for the Assignment of Genomic Fragments." *Environmental Microbiology* 6 (9) (September): 938–47.

Vernadsky, W I. 1945. "The Biosphere and the No{ö}sphere." *American Scientist* 33 (1): xxii–12.

Walsh, David a, Eric Bapteste, Masahiro Kamekura, and W Ford Doolittle. 2004. "Evolution of the RNA Polymerase B9 Subunit Gene (rpoB') in Halobacteriales: a Complementary Molecular Marker to the SSU rRNA Gene." *Molecular Biology and Evolution* 21 (12) (December): 2340–51.

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole. 2007. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73 (16) (August): 5261–7.

Watanabe, Keiji, Nobuyuki Komatsu, Yuichi Ishii, and Masami Negishi. 2009. “Effective Isolation of Bacterioplankton Genus Polynucleobacter from Freshwater Environments Grown on Photochemically Degraded Dissolved Organic Matter.” *FEMS Microbiology Ecology* 67: 57–68.

Wayne, L. G., D. J. Brenner, R. R. Colwell, P. a. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, et al. 1987. “Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics.” *International Journal of Systematic Bacteriology* 37 (4) (October 1): 463–464.

Wilson, Micheal C, and Jörn Piel. 2013. “Metagenomic Approaches for Exploiting Uncultivated Bacteria as a Resource for Novel Biosynthetic Enzymology.” *Chemistry & Biology* 20 (5) (May 23): 636–47.

Woese, Carl R. 1987. “Bacterial Evolution.” *Microbiological Reviews* 51 (2): 221–271.

Wright, Erik S, L Safak Yilmaz, and Daniel R Noguera. 2012. “DECIPHER, a Search-based Approach to Chimera Identification for 16S rRNA Sequences.” *Applied and Environmental Microbiology* 78 (3) (February): 717–25.

Wrighton, Kelly C, Brian C Thomas, Itai Sharon, Christopher S Miller, Cindy J Castelle, Nathan C VerBerkmoes, Michael J Wilkins, et al. 2012. “Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla.” *Science (New York, N.Y.)* 337 (6102) (September 28): 1661–5.

Wu, Dongying, Philip Hugenholtz, Konstantinos Mavromatis, Rüdiger Pukall, Eileen Dalin, Natalia N Ivanova, Victor Kunin, et al. 2009. “A Phylogeny-driven Genomic Encyclopaedia of Bacteria and Archaea.” *Nature* 462 (7276) (December 24): 1056–60.

Wu, Qinglong L, Gabriel Zwart, Michael Schauer, Miranda P Kamst-van Agterveld, and Martin W Hahn. 2006. “Bacterioplankton Community Composition Along a Salinity Gradient of Sixteen High-mountain Lakes Located on the Tibetan Plateau, China.” *Applied and Environmental Microbiology* 72 (8) (August): 5478–85.

Yannarell, Anthony C, and Eric W Triplett. 2005. “Geographic and Environmental Sources of Variation in Lake Bacterial Community Composition †.” *Society* 71 (1): 227–239.

Zerbino, Daniel R, and Ewan Birney. 2008. “Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs.” *Genome Research* 18 (May): 821–829.

Chapter 2: Temporal Scales of Community Diversity and Variation within and across Seven Humic Bog Lakes

Authors:

Benjamin Crary, Ashley Shade, Jack Gilbert, J. Gregory Caporaso, Rob Knight, and Katherine D. McMahon

Prologue:

Chapter 2 is a draft manuscript that will be submitted for publications. The target journal is ISMEJ. Author contributions are as follows:

Benjamin Crary: Wrote original draft of manuscript. Sequence analysis and statistical analyses.

Ashley Shade: Sample Collection.

Jack Gilbert: Read Library Construction.

J. Gregory Caporaso: Read Library Construction.

Katherine D. McMahon: Principle Investigator

Abstract

Bog lakes are reservoirs for large quantities of terrestrially derived organic matter, and the bacterial communities that inhabit them are the primary mineralizers within the ecosystem. To ultimately better understand how the bacterial communities may respond to regional and global changes, we seek to characterize the dynamics of community structure on short and long time scales. Here we amplified and sequenced the V4 hypervariable region of the 16S rRNA gene from a rigorously sampled historical time-series collected from seven different humic bog lakes. Thus, the bacterial communities were characterized for over 750 samples, with samples covering four calendar years. This unique and extensive dataset allowed us to assess community composition trends at multiple temporal and spatial scales. Rarefaction curves from each lake revealed that the bulk of microbial diversity within these lakes could be captured with approximately 25 samples taken throughout a year, but less abundant operational taxonomic units contributed to measurable differences in annual diversity. Overall diversity appeared to be driven by local environmental factors, as epilimnion and hypolimnion diversity within a particular lake and year were more comparable than across other spatial or temporal scales. The rate of change within bacterial communities was highest in the early summer, during dissolved nutrient and inorganic carbon depletion. Mixing events also created abrupt periods with high rates of change, and these changes were largely biological responses as opposed to apparent changes due to physical mixing.

Introduction

The biogeochemical processes occurring within lakes are strongly linked to their surrounding terrestrial landscapes. Since lakes serve as sentinels to regional and global climate changes (Adrian et al. 2009; Williamson et al. 2008), it is paramount that we understand the diversity and function of microorganisms that underpin the recycling of carbon and nutrients back into the food web. Understanding microbial community variation with time will enable more accurate predictions of

ecosystem responses to stochastic, regional, and global changes (Sugihara 1995). The merit to studying aquatic ecosystems has been outlined well, but the characteristic scales at which bacterial community composition needs to be studied are poorly defined (Lindström and Langenheder 2012).

With this objective in mind, global efforts are be made to document the microbial landscape (Gilbert et al. 2010; Caporaso et al. 2011), and often focus is given to microbial community dynamics. Annual patterns have shown that samples taken from bacterial communities from the same year can be more similar than those taken from a different year in some lakes, but not others (Jones et al. 2012). Long-term temporal studies have demonstrated that similarity in lake community structures decay slowly with time (Gonzalez et al. 2012), yet repeatable seasonal dynamics correlating to phytoplankton succession (Angela D Kent et al. 2007), predation (Pernthaler 2005), or resource availability (Wu and Hahn 2006).

Here we investigated the temporal scales of bacterial diversity within temperate bog lakes, and address the scale of which it is appropriate to assess the diversity of these dynamic and heterogeneous ecosystems. All of the study sites are humic bog lakes, which are reported to have a high synchrony of abiotic factors compared to other aquatic ecosystems (Järvinen et al. 2002). The close proximity of the study sites limits heterogeneity in regional-scale drivers, which would otherwise differentially influence abiotic and biotic synchrony across ecosystems (Magnuson et. al., 2005). Because of our unique dataset of weekly samples from seven different lakes during the ice-off period over four years, we are able to define the variation in samples on weekly, seasonal, and annual scales, as well as within- and across-lakes.

Finally, we examined this dataset with the intention of understanding not the community of a specific lake, but rather with the goal of understanding all lakes with similar ecosystem characteristics. The lakes in this study represent characteristic humic lakes and are surrounded by *Sphagnum* mosses, which have accumulation rates of 17-38 gCm^2/year (Kratz and Dewitt 1986). There is a relatively small

amount of productivity in the bog mats surrounding these lakes, yet the lakes themselves have prevalent primary producers (Kent et al. 2004) while also significant seepage of humic matter and DOC from the mats (Jones et al. 2009). Understanding the temporal trends in bacterial community structure will help researchers predict how decomposition and other ecosystems services may be altered given regional or global disturbances, such as climate change.

Methods

Study sites

Samples were collected from seven shallow bog lakes located in Vilas County, Wisconsin, USA (Table 2-1). Each lake is a humic lake and was chosen because they represent a range of mixing regimes. Sphagnum mats surround each lake, and leachate through these mats contributes to a high humic material content and dark stain. These lakes receive no hydrologic inputs from streams or rivers, and are isolated from human perturbations. Several of the lakes are sites under intense study as part of the North Temperate Lakes Long Term Ecological Research (NTL-LTER) site (http://lter.limnology.wisc.edu/).

Sample collection

Water samples were collected from the epilimnion and hypolimnion from the deepest part of the pelagic zone using an integrated water column sampler constructed from PVC pipe. Samples were taken approximately every week during the ice-free season, as described previously (Yannarell et al. 2003; Shade et al. 2008). Several time periods exist for select lakes in which samples were taken on consecutive days. Exact sample dates are shown in Supplemental Figure S2-1.

For Crystal Bog, which was not always stratified during sampling periods, the top 1 meter was collected and stored as the epilimnion sample, and the bottom 1 meter was collected and stored as the

hypolimnion sample. For the remaining lakes, temperature profiles were used on site to gauge the depth of the variable thermocline before sample collection. Bacteria were recovered by filtration on 0.2-mm polyethersulfone filters (Pall-Supor-200) without prefiltration promptly following sample collection. Filters were stored at -80 degrees C until sample processing.

Sample processing

Total DNA was extracted from filters using FastDNA Spin Kit (FastDNA® SPIN KIT, MP Biomedicals) with slight modifications to increase yield (Fisher and Triplett 1999). PCR amplification was performed on the V4 hypervariable region of the 16S rRNA gene using 515f and 806r primer set, followed by sequencing on the Illumina HiSeq platform(Caporaso et al. 2011). Sequencing reads were denoised and demultiplexed in QIIME with default settings (Caporaso et al. 2010).

OTU assignments

All sequences were processed with Mothur, using the Schloss SOP with minor modifications (Schloss et al. 2009). Samples were rarified to 5000 reads prior to clustering into OTUs with a cutoff of 0.10 to target 98% similarity with the average neighbor linkage algorithm.

Statistical analysis

Rarefaction curves were generated in Mothur using a trimmed dataset. The trimmed dataset was comprised of the most abundant 90% of data from the untrimmed dataset. This was done to eliminate singletons and other rare taxa that appeared in only a small number of samples. To generate rarefaction curves, samples from each lake-layer were pooled together, and samples were chosen in a random

sequence. Reported values of Shannon dissimilarity and Chao1 diversity were calculated using an untrimmed dataset.

Rate of change (RoC) was estimated by first calculating Bray-Curtis dissimilarity (Legendre and Legendre, 1998) between all samples based on the 90% trimmed OTU table. The Bray-Curtis dissimilarity was calculated with the following formula:

$$D_{BC} = 1 - 2\,\frac{\sum \min\left(\mathrm{S_{A,i}}, S_{B,i}\right)}{\sum S_{A,i} + \sum S_{B,i}}$$

where $S_{A,i}$ is the number of individuals in the *i*th OTU of community A and $S_{B,i}$ is the number of individuals in the OTU of community B.

Next, principle coordinates were found for all samples belonging to the same lake, layer, and year. The two major axes were used to find the 'distance' between two consecutive time points, and this distance was divided by time between the two samples. The following equations demonstrate the method used:

$$Distance_{ji} = \sqrt{\left(S_j axis1 - S_i axis1\right)^2 + \left(S_j axis2 - S_i axis2\right)^2}$$

$$RoC_j = \frac{Distance}{DayS_j - DayS_i}$$

where $Distance_{ji}$ is the distance between two consecutive samples points, *i* and *j*, on the two primary principle coordinate axes. $S_j axis1$, $S_i axis1$, $S_j axis2$, and $S_i axis2$ are the coordinates of the samples *j* and *i* from the primary axes of the principle coordinate. RoC_j is the rate of change for sample *j*, in reference to the previous time point sample *i*. $DayS_j$-$DayS_i$ is the amount of time between samples *j* and *i*.

Statistical analyses were performed with the aid of Mothur or R (R Development Core Team, 2012). The ggplot2 graphical package was used for figure generation (Wickham, 2009).

Results

Rarefaction Curve Saturation

We first asked whether the samples collected from each lake-layer adequately captured the overall diversity of that environment. After trimming out the least abundant 10% of reads in the dataset, the slope of the rarefaction curve for each lake-layer became less than 45 degrees after random sampling of approximately 20 samples (Figure 2-1). After approximately 25 samples, visual evidence suggests that each curve was approaching a plateau.

Diversity reached an apex at a similar rate for each lake, yet the lakes still varied in their overall richness. Deeper lakes, which also had stronger stratification periods, had an overall higher richness than those lakes that were shallower and had weaker stratification. The differences between these lakes were apparent on the rarefaction curve after five random samples. Differences between the shallower lakes were not clear until after at least ten samples were compared.

The diversity was more consistent between layers in a single lake than for similar layers across lakes. In five of the seven lakes, the epilimnion and hypolimnion rarefaction curves were separated by very few OTUs. In Forestry Bog and West Sparkling Bog, the hypolimnion had 10% and 4% more observed OTUs, respectively.

Annual Patterns of Diversity

We calculated the average and variance of the Shannon's diversity index from all samples in the same lake, layer, and year to understand the range in diversity across long (multi-year) temporal scales. The results demonstrate that overall Shannon's diversity index for temperate (boreal?) bog lakes hovers near 4 (Figure 2-2). While each lake followed this general trend, the diversity of each lake was not predictable for a given year. For each lake layer that was sampled for multiple years, there was at least

one sample year in which the mean diversity was outside the variance for the remaining sample years. For example, the hypolimnion of MA in 2007 was statistically different from the hypolimnion of MA in 2005, 2008, and 2009 in respect to Shannon's diversity.

Other lakes followed this pattern as well, but with larger differences in the means. The hypolimnion samples seemed to be especially variable, particularly for MA, SSB, and TB. The variation within epilimnion samples from these three lakes was also more pronounced.

The diversity from both layers within the same lake often followed the same trend through time. That is, if the diversity in the epilimnion increased from one year to the next, the hypolimnion was likely to follow this trend. There were 11 sample sets (samples coming from the same year for a specific lake) in which there was another sample set for the same lake at a subsequent sample year. In 73% percent of these sets, the epilimnion and hypolimnion followed the same directional diversity trend.

Comparing all lakes, SSB had the most variation in diversity across years. The mean diversity of both layers was statistically different for each of the years sampled. Alternatively, NSB was fairly consistent in both layers across years. The within-year variation for NSB also appeared to be consistently among the least variable.

Seasonal and Sub-Seasonal Rates of Change

We estimated the community RoC by quantifying the distance between sample points on the major two axes of a PCoA analysis plot. After placing the samples into weekly bins, we observed the average RoC for each week (Figure 2-3). Overall, both the epilimnion and the hypolimnion followed the same seasonal trends and had similar magnitudes in the RoC and variation. The average RoC was the highest and most variable during the early summer period (weeks 23-34). The highest concentration of outliers occurred during this season as well. The late summer period (weeks 35-43) appeared to be the

most stable in terms of, both, the RoC and the variability of change across the lakes. Again, this was observed for both the epilimnion and hypolimnion layers.

With fewer time points in the spring and autumn seasons, it was more difficult to assess the true seasonal RoC. Though fewer weeks were examined, the samples from week 19 were especially interesting when noting that 83% of the sample set from the epilimnion and 100% of the samples from the hypolimnion (epilimnion total n=6, hypolimnion total n=5) were taken on consecutive days from NSB. This week had a considerably high RoC and a high amount of variation in the epilimnion, especially for samples that were taken daily. Interestingly, the outlier from week 19 did not come from the single sample from TB, but rather one of the NSB samples. The RoC in the TB sample was slightly below the average (RoC=0.01) for that week. For the six samples taken from the hypolimnion in week 19, the RoC was less than what was observed in the epilimnion, but the RoC from this week was still greater than all but one late summer hypolimnion sample (week 42).

In a similar fashion the samples taken during week 46 were all consecutive days sampled from TB in 2007. There was extreme change in the epilimnion and hypolimnion in week 46, with a high degree of variation. The high RoC and variation in rate came to an abrupt halt beginning in week 47. Excluding weeks 19, 46, and 47, the RoC from each week was estimated with samples taken from multiple lakes.

Variation at the Scale of Days

The samples taken during weeks 45 to 47 from TB during 2007 were selected to observe change in the two layers during fall mixing. Fall mixing was chosen as a study period to investigate the magnitude of community change upon temperature and dissolved oxygen perturbations, since these two factors often drive bacterial community composition (Lindstrom et al. 2005). By constructing temperature and dissolved oxygen profiles for these dates, we could confirm that these samples captured the tail end of fall turnover (Figure 2-4). The epilimnion thermal layer had begun to gradually expand and lower in

temperature approximately two weeks prior to these samples, but the samples taken during weeks 45 to 47 captured the water column becoming isothermal on November 5th. Additionally, these samples captured the rapid oxygenation of the water column.

By creating a PCoA with only these fall overturn, we were able to assess how similar the epilimnion and hypolimnion samples became during mixing (Figure 2-5A). We observed that the samples from the two layers were distinct from one another on November 5th, but the samples from each layer had converged by November 8th. No subsequent divergence occurred in the remaining nine samples. The estimated richness of the samples from each layer followed similar patterns (Figure 2-5B). A linear decline in richness was apparent through November 11th for both layers (R^2=0.52), but no positive or negative linear relationship was observed for samples collected after November 11th (R^2<0.1).

Discussion

Overall Patterns of Diversity

Trends in diversity could be observed at different sampling effort thresholds for the lakes surveyed (Figure 2-1). The three deepest lakes that stratify the strongest throughout the summer could be distinguished from the shallower lakes after approximately five random samples on a rarefaction curve. Visual distinction between the shallower lakes was not apparent until approximately 12 random samples. These approximations, though made from visual assessment of a rarefaction curve, begin to illustrate the depth of sampling necessary to adequately compare across similar, but yet distinct, environments. While differences between lakes, in terms of overall diversity, could be observed with relatively few samples, at least 25 samples were needed to approach the plateau.

It was interesting to observe that MA was among the most diverse lakes. Typically, meromictic lakes would be considered a stable ecosystem in comparison to polymictic and even dimictic lakes. If

these lakes were to be compared along a gradient of mixing disturbances, MA would be the least disturbed of these lakes, and the bacterial community would be among the least diverse according to the intermediate hypothesis theory (Connell 1978). The remaining lakes, however, could be compared favorably with this hypothesis. SSB and TB, which are both dimictic (intermediate number of disturbances), were more diverse than the remaining polymictic lakes (high number of disturbances). It appears as though these lakes have strong enough environmental variation as to support to this hypothesis.

Although the intermediate disturbance hypothesis would suggest that MA would have the lowest diversity, other ecological factors such as strong physical and chemical gradients within MA may be more important in determining overall diversity. The maximum depth of MA is twice the next deepest lake, and is characterized by a well-developed redox gradient, which may provide increased niche space. The chemical gradients that contribute to keeping the lake permanently stratified and subsequently less disturbed seasonally provide conditions that are not present in the other lakes. Further, MA is hydrologically connected to another small lake (Rose Lake) during the spring season when water levels are still high from snowmelt. Rose Lake has a unique physiochemical nature, yet this hydrologic connection allows immigration and emigration between these two systems, which likely impacts the measured diversity. Thus the meromictic nature and unique hydrology of MA suggests that the intermediate disturbance hypothesis may only explain changes in diversity within ecosystem types but is less robust in explaining changes in diversity across ecosystems.

Perhaps unsurprisingly, the 25 samples needed to adequately gauge the bulk of diversity within a lake was near the number of samples that were collected throughout a year at a twice-monthly frequency. For those lakes with only one sample year (WSB and FB), it did appear as though plateaus were forming, however the impact of additional sampling years may certainly have increased the observed diversity. Since the rarefaction curves were constructed using the most abundant 90% of the data, these findings imply that the majority of the diversity within a lake ecosystem can be identified by sampling a particular

lake approximately 25 times over the course of a year. It is possible that periods of rapid succession occur on a scale finer than our weekly sampling regime and that we underestimated diversity by not characterizing communities as quickly as they change. By sampling over multiple years, however, we minimized the effect of missing rapid changes on fine temporal scales. Though, if the rare OTUs are of interest, then a single year of sampling may not be adequate to accurately estimate true richness.

The inclusion of more rare OTUs into this analysis would have altered our assessment of how adequate the datasets were to make ecological inferences. While we only included the top 90% of the data into rarefaction curves, over 500 OTUs were observed for each lake. This number of OTUs is considerably higher than the number achieved with community fingerprinting techniques (Shade et al. 2008),and we argue that by disregarding the least abundant 10% of data we are removing OTUs and taxa that are not persistently present in time-series data.

Annual Patterns of Diversity

When separated out by year, we observed that the study sites had Shannon diversity near 4 for any give year, yet there was not annual stability. Because diversity peaked for lakes in different years, we were able to infer that there were still localized characteristics that drove diversity, even for lakes sampled from the same climactic region. For example, the peak diversity in TB occurred in 2005, the peak diversity in MA occurred in 2007, and the peak diversity in SSB occurred in 2009. Without assessing the annual trends of other ecosystems, it is difficult to speculate whether these local drivers were stochastic or deterministic. That is, was there a local event, such as a phytoplankton bloom of a rare species, which drove changes in the community structure from year to year, or were the observed spikes in diversity of MA, SSB, and TB within the normal range for bog lakes and caused by seemingly random interactions.

Another interesting facet of the annual patterns was the relationship between the epilimnion and hypolimnion samples for a given lake and year. Though the two layers are often considered to be distinct environments with distinct community compositions and patterns (Ochs, et al, 1995; Shade et al. 2010),

we observed that most often (73%), the diversity change in both layers was from year to year was in the same direction. The strong correlation across layers from the same lake was also evident in the total diversity (Figure 2-1) and with an AMOVA analysis comparing compositions (Supplemental Table S2-1).

These results indicate that diversity is driven more by the local factors common to both layers that can vary year to year than the features that define the epilimnion and the hypolimnion. Phytoplankton could have an effect on both epilimnion and hypolimnion layers, as their detritus will sink into the hypolimnion, providing a common source of organic matter to both layers. Additionally, pH in the two layers will be locally influenced by the humic acid seepage into the lake, inorganic carbon uptake by photosynthetic organisms, and respiration. These layers are linked in other ways, thus it appears that by sampling a specific depth or layer, information can be inferred about the community dynamics or diversity at other depths throughout the lake. The results in this study support that concept, yet the communities observed in either layer are in almost all cases compositionally distinct, and layer or depth specific processes must be sampled for appropriately.

Seasonal Patterns of Change

Understanding community composition and the factors that drive it can help predict function and responses to change (Chapin et al. 2000), and recent success has been had in terms of predicting the temporal make up of a bacterial community (Fuhrman et al. 2006; Gilbert et al. 2012). While these are great advances, there is still little known about the RoC within most ecosystems, particularly those with strong seasonal changes within a year. The rate of variability or succession in communities is most often lake-specific because of the dissimilarities in environmental conditions (Rusak et al. 1999; Baines et al. 2000).

Here, we were able to identify seasonal patterns in the successional pace of lakes with overall similar characteristics. In the bog lakes studied, the average rate of change was highest throughout the

early summer in both the epilimnion and hypolimnion (Figure 2-3). Considering that lakes are undergoing thermal stratification and rapid shifts in phytoplankton community composition (Kent et al. 2006; Kent et al. 2007), among other physical and chemical changes during this period (Supplemental Figures S2-2 and S2-3), these results suggest that environmental factors are a major factor in governing the rate at which bacteria communities assemble across seasons.

With our findings, we stress the necessity to focus sampling efforts during early stratification periods, and the subsequent weeks, in order to adequately document seasonal diversity. The rapid depletion of nutrients and inorganic carbon during early summer likely causes a large bottom up and top down influence on bacteria. Previous studies have identified distinct and repeated seasonal communities (Kent et al. 2006; Shade 2008), but few have attempted to quantify the rate at which the communities are changing in a given season.

We also note the occurrence of rather brief (~1 week) periods in which major community shifts may occur given large ecosystem disturbances such as lake mixing. In particular, the fall mixing period was considered in more detail on a daily scale in TB (Figures 2-4 and 2-5), but it is worth pointing to the magnitudes of change between mixing events and nutrient depletion. In the epilimnion, the RoC throughout the early summer matched the RoC of the physical mixing event in the fall (Figure 2-3). This suggests that the top down and bottom up pressures resulting from predation and nutrient depletion have similar magnitudes of impact as the environmental changes caused by physical mixing.

Daily Succession

We addressed the issue of community convergence on a daily scale by targeting a time frame during which large community compositional change was suspected to be occurring. At the onset of mixing, the thermocline gradually lowers through the water column, the water column becomes

oxygenated, and the sediments release chemically bound nutrients. This presumably has large impacts on the biological processes as the epilimnion and hypolimnion merge.

Here we captured the communities while they were still distinct, and observed the manner in which they converged during fall mixing (Figure 2-4). Using PCoA to visualize the sampled community structure (Figure 2-5A), it appeared that the hypolimnion and epilimnion partially underwent the same trajectory, except that the changes in the epilimnion occurred a couple of days ahead of the hypolimnion. We observed that the first three hypolimnion samples captured directional compositional changes that converged to a state observed in the November 5th epilimnion sample. When the hypolimnion reached that community composition on November 7th, it followed the trajectory that the epilimnion had already begun.

We estimated the chao1 richness for these samples as a way to gauge whether the community composition shifts were in fact biological responses, or simply a phenomenon of physical mixing. If the responses were physical mixing, we would have expected to observe the richness increase because we would be increasing the number of rare taxa in the sample. This was confirmed by subsampling two *in-silico* communities comprised of 50 OTUs with 250 observations each (Shannon's index = 2.8) for 250 observations 1000 times (Supplemental Material). Instead, we observed a linear decrease in richness until Nov 11th approximately the date at which any distinction between epilimnion and hypolimnion samples was absent. Thus, we conclude that the community composition changes during mixing were in large due to biological factors (e.g. competition).

Although biological responses were observed during mixing, it is interesting to note that the chao1 richness estimates of the epilimnion and hypolimnion samples were not correlated with each other after November 12th. The Bray-Curtis dissimilarity metrics revealed that the overall community structures were very similar (Figure 2-5A), but the hypolimnion samples on the 15th, 16th and 17th were considerably more diverse (Figure 2-5B). This likely suggests that although the lake was essentially isothermal with an

even dissolved oxygen concentration throughout the water column, there are other factors (such as nutrients released from the sediments after the water column became oxygenated) that allow a more diverse community to thrive at lower depths during mixing.

There is much to gain from understanding how a community responds to disturbances, and on what temporal scales. Future studies can use these particular results in multiple ways. For example, researchers that are interested in the community function and composition during a disturbance should be more capable of selecting a sampling regime that is relevant to their question. Here we saw that there were several days of rapid change followed by very little change for several days. Alternatively, these results suggest that by observing microbial community dynamics, researchers should be able to predict when disturbances occurred.

Concluding Remarks

From this temporal analysis, we gain a stronger understanding of how aquatic communities vary in their structure through time. This work, which represents one of the most highly sampled time series studies of microbial community composition to date, provides insight into community dynamics on relevant ecological scales. Sampling efforts and experiments can be designed more optimally because of this work. Here we identify that diversity across years is moderately consistent for similar ecosystems, but the diversity in a given system is still unpredictable on an annual basis. We also identified early summer as a time in which the rate of change in elevated relative the rest of the seasons. We also observed, however, that community composition can rapidly change given a disturbance (such as water-column mixing) before abruptly maintaining a particular composition outside of the early summer period.

Studying how bacterial communities vary through time will bring researchers and engineers closer to constructing effective ecological models that include microbial dynamics in a fashion that is more informative that the “black box” (Tiedje et al. 1999; Shade et al. 2009). Lakes are considered

sentinels of change, thus understanding the microbial communities that underpin these ecosystems will help in the prediction of ecosystem responses to stochastic, regional, and global changes (Sugihara 1995; Williamson et al. 2008; Adrian et al. 2009). We hope that future studies can elaborate further on the temporal patterns exhibited by microorganisms.

Jones, Stuart E, Tracey a Cadkin, Ryan J Newton, and Katherine D McMahon. 2012. "Spatial and Temporal Scales of Aquatic Bacterial Beta Diversity." *Frontiers in Microbiology* 3 (August) (January): 318.

Jones, Stuart E, Ryan J Newton, and Katherine D McMahon. 2009. "Evidence for Structuring of Bacterial Community Composition by Organic Carbon Source in Temperate Lakes." *Environmental Microbiology* 11 (9) (September): 2463–72.

Järvinen, M, M Rask, J Ruuhijärvi, and L Arvola. 2002. "Temporal Coherence in Water Temperature and Chemistry Under the Ice of Boreal Lakes (Finland)." *Water Research* 36 (16) (September): 3949–56.

Kent, A D, S E Jones, A C Yannarell, J M Graham, G H Lauster, T K Kratz, and E W Triplett. 2004. "Annual Patterns in Bacterioplankton Community Variability in a Humic Lake." *Microbial Ecology* 48 (4) (November): 550–60.

Kent, Angela D, Stuart E Jones, George H Lauster, James M Graham, Ryan J Newton, and Katherine D McMahon. 2006. "Experimental Manipulations of Microbial Food Web Interactions in a Humic Lake: Shifting Biological Drivers of Bacterial Community Structure." *Environmental Microbiology* 8 (8) (August): 1448–59.

Kent, Angela D, Anthony C Yannarell, James a Rusak, Eric W Triplett, and Katherine D McMahon. 2007. "Synchrony in Aquatic Microbial Community Dynamics." *The ISME Journal* 1 (1) (May): 38–47.

Kratz, Timothy K., and Calvin B Dewitt. 1986. "Internal Factors Controlling Peatland-Lake Ecosystem Development." *Ecology Society of America* 67 (1): 100–107.

Legendre P, Legendre L. (1998). Numerical Ecology. Elsevier Science: BV, Amsterdam.

Lindstrom, Eva S., Miranda P. Kamst-Van Agterveld, and Gabriel Zwart. 2005. "Distribution of Typical Freshwater Bacterial Groups Is Associated with pH , Temperature , and Lake Water Retention Time." *Applied and Environmental Microbiology* 71 (12): 8201–8206.

Lindström, Eva S, and Silke Langenheder. 2012. "Local and Regional Factors Influencing Bacterial Community Assembly." *Environmental Microbiology Reports* 4 (1) (February): 1–9.

Magnuson JJ, Benson BJ, Lenters JD, Robertson DM. (2005). Coherent dynamics among lakes. In: Magnuson JJ, Kratz TK, Benson BJ (eds). Long Term Dynamics of Lakes in the Landscape. Oxford Press: Oxford, pp 89–106.

Ochs, C. A., J.J. Cole & G.E. Likens. 1995. Population dynamics of bacterioplankton in an oligotrophic lake. J. Plankton Res., 17: 365-391.

Pernthaler, Jakob. 2005. "Predation on Prokaryotes in the Water Column and Its Ecological Implications." *Nature Reviews. Microbiology* 3 (7) (July): 537–46.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

References

Adrian, Rita, Catherine M O'Reilly, Horacio Zagarese, Stephen B Baines, Dag O Hessen, Wendel Keller, David M Livingstone, et al. 2009. "Lakes as Sentinels of Climate Change." *Limnology and Oceanography* 54 (6) (November): 2283–2297.

Baines, Stephen B., Katherine E. Webster, Timothy K. Kratz, Stephen R. Carpenter, and John J. Magnuson. 2000. "Syncrhonous behavior of temperature, calcium, and chlorophyll in lakes of northern Wisconsin." *Ecology* 81 (3): 815–825.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High- Throughput Community Sequencing Data I." *Nature Publishing Group* 7 (5): 335–336.

Caporaso, J Gregory, Christian L Lauber, William a Walters, Donna Berg-Lyons, Catherine a Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. 2011. "Global Patterns of 16S rRNA Diversity at a Depth of Millions of Sequences Per Sample." *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl (March 15): 4516–4522.

Chapin, F S, E S Zavaleta, V T Eviner, R L Naylor, P M Vitousek, H L Reynolds, D U Hooper, et al. 2000. "Consequences of Changing Biodiversity." *Nature* 405 (6783) (May 11): 234–42.

Connell, Joseph H. 1978. "Diversity in Tropical Rain Forests and Coral Reefs." *Science* 199 (4335): 1302–1310.

Fisher, Madeline M, and Eric W Triplett. 1999. "Automated Approach for Ribosomal Intergenic Spacer Analysis of Microbial Diversity and Its Application to Freshwater Bacterial Communities." *Applied and Environmental Microbiology* 65 (10): 4630–4636.

Fuhrman, Jed a, Ian Hewson, Michael S Schwalbach, Joshua a Steele, Mark V Brown, and Shahid Naeem. 2006. "Annually Reoccurring Bacterial Communities Are Predictable from Ocean Conditions." *Proceedings of the National Academy of Sciences of the United States of America* 103 (35) (August 29): 13104–9.

Gilbert, Jack A, Folker Meyer, Janet Jansson, Jeff Gordon, Norman Pace, James Tiedje, Ruth Ley, et al. 2010. "The Earth Microbiome Project : Meeting Report of the ' 1 St EMP Meeting on Sample Selection and Acquisition ' at Argonne National Laboratory October 6 Th 2010 ." *Standards in Genomic Sciences* 3 (3): 249–253.

Gilbert, Jack A., Joshua A. Steele, J Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, et al. 2012. "Defining Seasonal Marine Microbial Community Dynamics." *The ISME Journal* 6 (2) (February): 298–308.

Gonzalez, Antonio, Andrew King, Michael S Robeson, Sejin Song, Ashley Shade, Jessica L Metcalf, and Rob Knight. 2012. "Characterizing Microbial Communities Through Space and Time." *Current Opinion in Biotechnology* 23 (3) (June): 431–6.

Rusak, James A, Norman D Yan, Keith M Somers, and Donald J Mcqueen. 1999. “The Temporal Coherence of Zooplankton Population Abundances in Neighboring North-Temperate Lakes.” *The American Naturalist* 153 (1): 46–58.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan a Lesniewski, et al. 2009. “Introducing Mothur: Open-source, Platform-independent, Community-supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23) (December): 7537–41.

Shade, A, S E Jones, and K D McMahon. 2008. “The Influence of Habitat Heterogeneity on Freshwater Bacterial Community Composition and Dynamics.” *Environmental Microbiology* 10 (4): 1057–1067.

Shade, A., C.-Y. Chiu, and K. D. McMahon. 2010. “Differential Bacterial Dynamics Promote Emergent Community Robustness to Lake Mixing: An Epilimnion to Hypolimnion Transplant Experiment.” *Environmental Microbiology* 12 (2) (February): 455–66.

Shade, Ashley, Cayelan C Carey, Emily Kara, Stefan Bertilsson, Katherine D McMahon, and Matthew C Smith. 2009. “Can the Black Box Be Cracked? The Augmentation of Microbial Ecology by High-resolution, Automated Sensing Technologies.” *The ISME Journal* 3 (8) (August): 881–8.

Shade, Ashley, Stuart E Jones, and Katherine D McMahon. 2008a. “The Influence of Habitat Heterogeneity on Freshwater Bacterial Community Composition and Dynamics.” *Environmental Microbiology* 10 (4) (April): 1057–67.

Sugihara, George. 1995. “From Out of the Blue.” *Nature* 378: 559–560.

Tiedje, James M., Stella Asuming-brempong, Klaus Nu, Terry L. Marsh, and Shannon J. Flynn. 1999. “Opening the Black Box of Soil Microbial Diversity.” *Applied Soil Ecology* 13: 109–122.

H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

Williamson, Craig E, Walter Dodds, Timothy K Kratz, and Margaret a Palmer. 2008. “Lakes and Streams as Sentinels of Environmental Change in Terrestrial and Atmospheric Processes.” *Frontiers in Ecology and the Environment* 6 (5) (June): 247–254.

Wu, Qinglong L, and Martin W Hahn. 2006. “High Predictability of the Seasonal Dynamics of a Species-like Polynucleobacter Population in a Freshwater Lake.” *Environmental Microbiology* 8 (9) (September): 1660–6.

Yannarell, A C, A D Kent, G H Lauster, T K Kratz, and E W Triplett. 2003. “Temporal Patterns in Bacterial Communities in Three Temperate Lakes of Different Trophic Status.” *Microbial Ecology* 46 (4) (November): 391–405.

Tables and Figure

Table 1-1: Limnological Characteristics of the Study Lakes

Lake	*Latitude (N)*	*Longitude (W)*	*Surface Area (Ha)*	*Max depth (m)*	*DOC (mg/l)*	*pH*	*Total P (µg/l)*	*Total N (µg/l)*	*Chl a (µg/l)*
Crystal Bog	46°00'26.8''N	89°36'22.5"W	0.56	2.5	9.5	5.1 (0.8)	22.3 (7)	629.0 (261)	22.4 (44.2)
Forestry Bog	46°02'51.4''N	89°39'04.8"W	0.13	2.5	10.4	5.5 (0.4)	27.2 (13)	830.5 (369)	28.1 (43.3)
*Mary Lake	46°15'02.6"N	89°54'00.7"W	1.20	21.5	26.4	5.5 (0.5)	30.0 (NA)	892.0 (NA)	NA
North Sparkling Bog	46°00'16.0"N	89°42'18.6"W	0.44	4.5	9.5	5.2 (1.2)	25.4 (11)	691.9 (206)	27.4 (71.3)
South Sparkling Bog	46°00'13.6"N	89°42'19.9"W	1.01	8.5	11.2	5.1 (0.9)	21.8 (20)	642.0 (161)	22.3 (34.7)
Trout Bog	46°02'27.5"N	89°41'09.6"W	1.23	7.9	28	4.8 (0.7)	31.6 (22)	815.2 (237)	8.17 (15.9)
West Sparkling Bog	46°00’16.6” N	89°42’33.5”W	1.5	4.6	NA	6.0	NA	NA	NA

Abbreviations: DOC, dissolved organic carbon; total N, total nitrogen; total P, total phosphorus; Chl a, chlorophyll a.

Numbers in parentheses represent the range of observed values (maximum value - minimum value).

Measurements were taken during 2003 (Kent et al., 2007)

*Measurements were taken during 2005 (Shade et al. 2008)

Table 1-2: Untrimmed Diversity Characteristics

Lake & Layer	Samples (Years)	Observed OTUs	Chao1	Inverse Simpson	Shannon
CBE	40 (2007,2009)	10428	42606	30.8	5.1
CBH	39 (2007,2009)	10762	43640	36.3	5.2
FBE	31 (2007)	8755	36794	16.3	4.4
FBH	32 (2007)	10842	38846	18.9	4.8
MAE	81 (2005,2007,2008,2009)	21501	79539	39.8	5.7
MAH	74 (2005,2007,2008,2009)	22571	80863	44.8	5.8
NSBE	71 (2007,2008,2009)	14468	48451	19.8	4.8
NSBH	77 (2007,2008,2009)	16637	57200	24.1	5.1
SSBE	49 (2007,2008,2009)	14868	50293	26.4	5.6
SSBH	58 (2007,2008,2009)	17315	56713	31.6	5.8
TBE	82 (2005,2007,2008,2009)	20818	78442	34.0	5.6
TBH	88 (2005,2007,2008,2009)	22833	86849	33.8	5.6
WSBE	28 (2007)	8020	35034	17.8	4.9
WSBH	27 (2007)	8747	36551	22.9	5.0

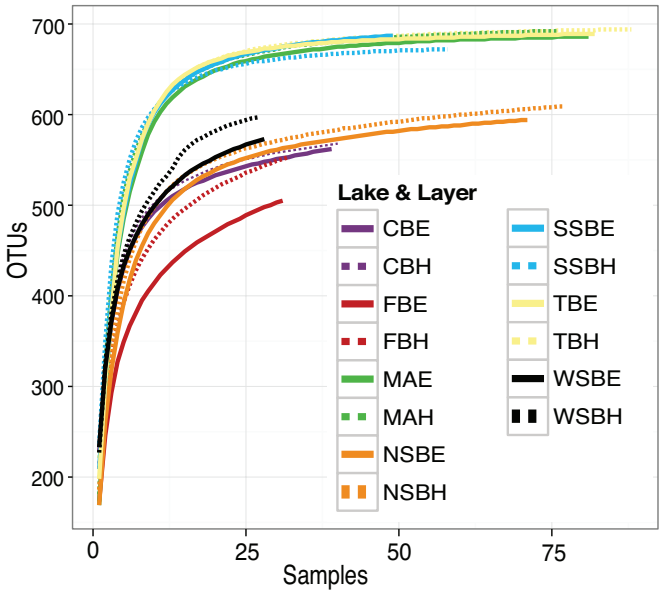

Figure 2-1: Lake and Layer Specific Rarefaction Curves. Before rarefactions curves were generated, the whole dataset was trimmed to only include the most abundant 90% of the data. Observations were taken from samples that were chosen in a random order. The curves began to slow at approximately 20 samples, and each curve appeared to be close to a plateau after 25 samples.

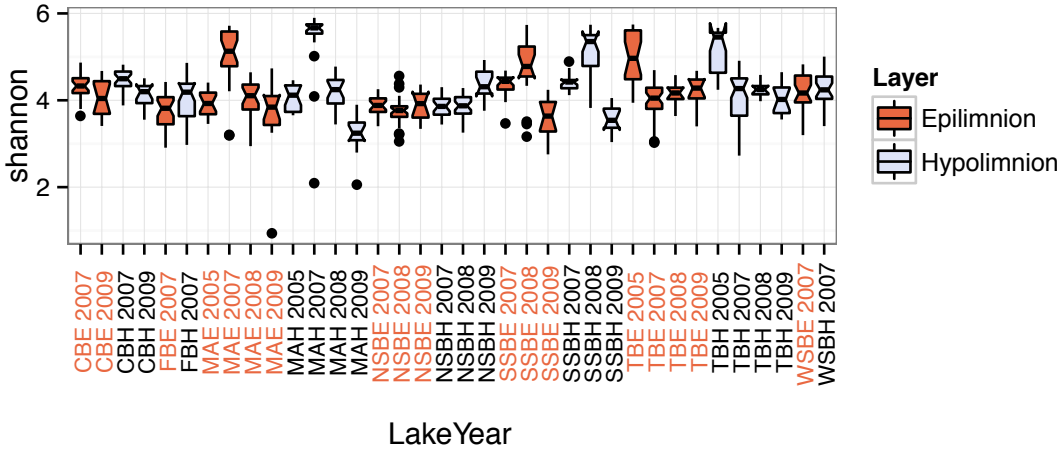

Figure 2-2: Shannon Diversity measured for each Lake, Layer, and Year. Diversity estimates were calculated with unfiltered OTU tables. Epilimnion estimates are colored in orange, hypolimnion estimates are colored in violet. The upper and lower box hinges refer to the first and third quartiles. Whisker length is 1.5* the interquartile range. Notches are drawn (1.58*the interquartile range)/(square root of n). Lines within the interquartile range represent the median. Differences were observed across years in the same lakes, though these differences were more pronounced for some lakes (i.e.: SSB) than others (CB). Similar trends were seen for epilimnion and hypolimnion layers across time within the same lake.

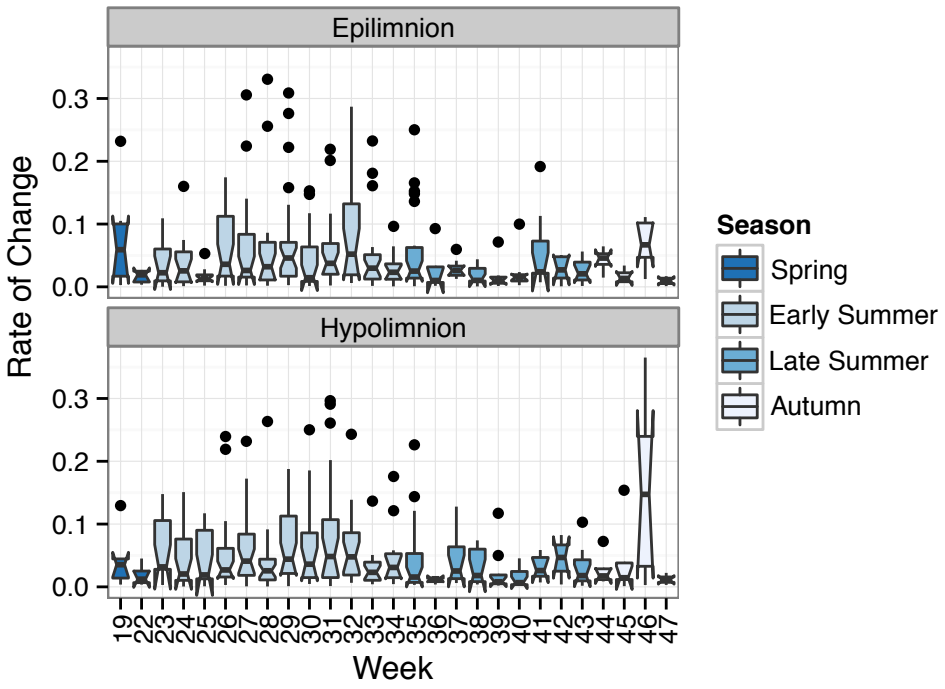

Figure 2-3: Average RoC through time in the epilimnion and hypolimnion. RoC was highest during early summer, and the lowest during late summer. Both layers followed similar seasonal trends and compared in terms of magnitude. Weekly periods that contained daily samples from the same lake and layer (Week 19, NSB; Weeks 46 and 47, TB) demonstrated that both it is possible to have very high and low rates of changes on a day–to-day period.

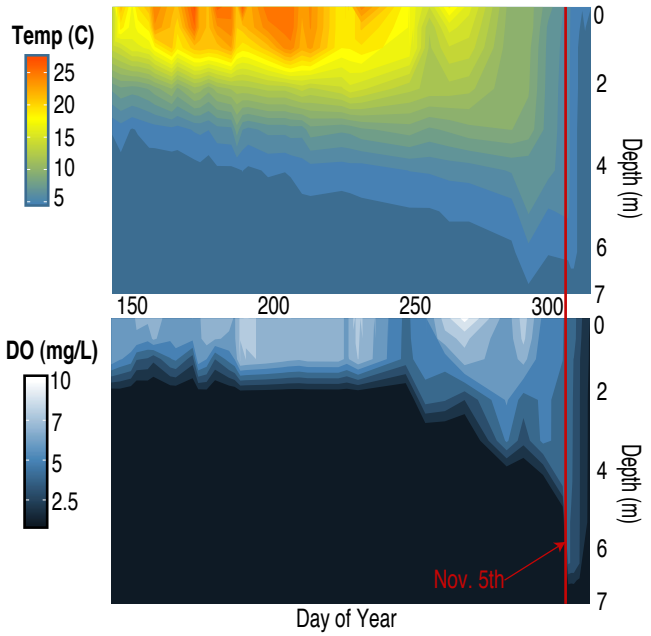

Figure2-4: Temperature and Dissolved Oxygen Profiles for Trout Bog in 2007. Temp, Temperature measured in degrees Celsius; DO, dissolved oxygen measured in mg/L. The red line is drawn on November 5th, when the first of 12 daily samples from the epilimnion and hypolimnion were taken in order to characterize community structure during fall turnover.

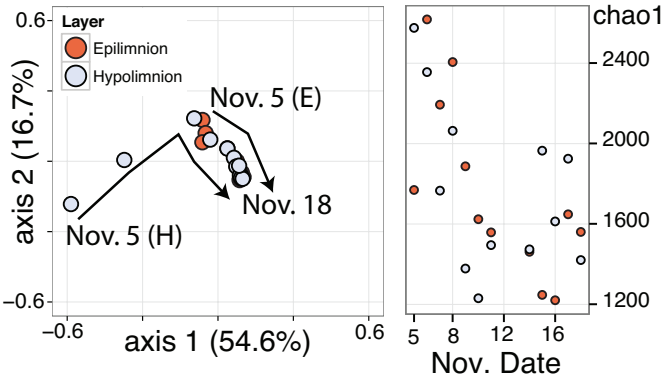

Figure 2-5: A) Principle coordinate analysis of daily samples taken from TB during fall mixing on November 5th, 2007 to November 18th, 2007. Dissimilarity was estimated via Bray-Curtis on an OTU table including only the top 90% of abundance. B) Chao1 estimates for the samples depicted in Figure 5A. Estimates were made on an OTU table including all OTU data. The communities from the epilimnion and hypolimnion converge to becoming more similar, while simultaneously becoming less rich. Change is rapid initially, but the change slows as the communities get more similar.

Supplemental Material

Supplemental Methods

Mixed Community Diversity Estimates:

Two identical *in-silico* communities comprised of 50 OTUs and 249 observations were randomly constructed. The Solver tool in Microsoft Excel was used to determine the abundance of each OTU while targeting a Shannon diversity index of 2.8. Several OTUs were 'seeded' with abundances near 5%-15% of the community to optimize the Solver solution towards the rank abundance observed in the lakes.

These two communities were then combined to simulate the mixing of the epilimnion and hypolimnion (100 total OTUs, 498 total observations). From this mixed community, 249 observations were randomly sampled to replicate environmental sampling. This subsampled community was then estimated for OTU richness. This was repeated 1000 times, and the average richness and variance was recorded.

Chemical Data

Data measurements for carbon, nitrogen, and phosphorous were taken biweekly on Trout as part of the North Temperate Lakes – Long Term Ecological Research program (http://lter.limnology.wisc.edu). Complete methodology can be found at http://lterquery.limnology.wisc.edu.

Supplemental Tables and Figures

Table S2-1: AMOVA statistics for pairwise comparisons of lake layers. Each cell reports an F-statistic followed by the p-value in parentheses.

	CBE	CBH	FBE	FBH	MAE	MAH	NSBE	NSBH	SSBE	SSBH	TBE	TBH	WSBE	WSBH
CBE	-													
CBH	0.73(.646)	-												
FBE	13.3(<0.001)	16.0(<0.001)	-											
FBH	11.5(<0.001)	13.2(<0.001)	1.9(0.07)	-										
MAE	6.5(<0.001)	7.3(<0.001)	1.8(<0.001)	8.4(<0.001)	-									
MAH	7.2(<0.001)	7.4(<0.001)	10.5(<0.001)	9.3(<0.001)	3.1(<0.001)	-								
NSBE	6.3(<0.001)	8.6(<0.001)	10.3(<0.001)	8.7(<0.001)	9.3(<0.001)	11.4(<0.001)	-							
NSBH	5.2(<0.001)	6.0(<0.001)	12.9(<0.001)	9.3(<0.001)	8.8(<0.001)	9.8(<0.001)	2.6(<0.001)	-						
SSBE	8.8(<0.001)	9.5(<0.001)	12.0(<0.001)	9.1(<0.001)	6.2(<0.001)	6.8(<0.001)	9.7(<0.001)	6.6(<0.001)	-					
SSBH	11.7(<0.001)	11.9(<0.001)	17.1(<0.001)	13.0(<0.001)	8.8(<0.001)	7.5(<0.001)	14.4(<0.001)	9.6(<0.001)	1.0(0.37)	-				
TBE	6.0(<0.001)	5.8(<0.001)	14.9(<0.001)	12.1(<0.001)	6.5(<0.001)	6.6(<0.001)	7.8(<0.001)	4.7(<0.001)	7.5(<0.001)	9.1(<0.001)	-			
TBH	8.2(<0.001)	8.2(<0.001)	20.0(<0.001)	15.4(<0.001)	11.8(<0.001)	9.6(<0.001)	13.0(<0.001)	8.0(<0.001)	9.4(<0.001)	10.4(<0.001)	2.9(0.006)	-		
WSBE	13.7(<0.001)	13.7(<0.001)	23.7(<0.001)	19.2(<0.001)	10.7(<0.001)	11.1(<0.001)	11.1(<0.001)	9.0(<0.001)	10.3(<0.001)	12.9(<0.001)	7.0(<0.001)	7.9(<0.001)	-	
WSBH	9.8(<0.001)	9.8(<0.001)	21.5(<0.001)	16.6(<0.001)	9.2(<0.001)	9.3(<0.001)	9.1(<0.001)	6.7(<0.001)	9.8(<0.001)	11.7(<0.001)	4.7(<0.001)	6.1(<0.001)	2.4(0.02)	-

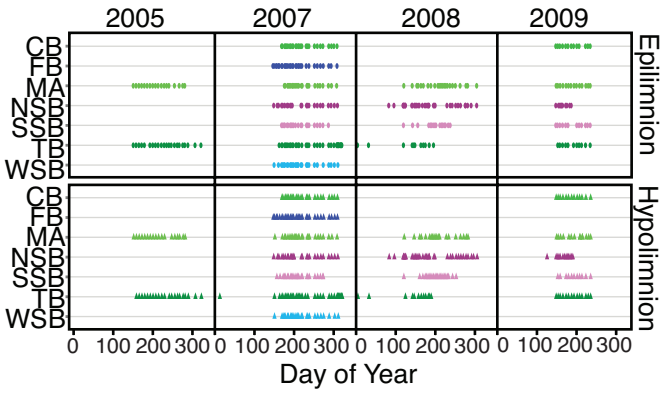

Figure S2-1: A four-year timeline illustrating the sample dates for each of the lakes sampled. Most sampling regimes were designed to be weekly over the ice-off season. The samples depicted represent only the samples that had at least 5000 quality reads prior to OTU clustering.

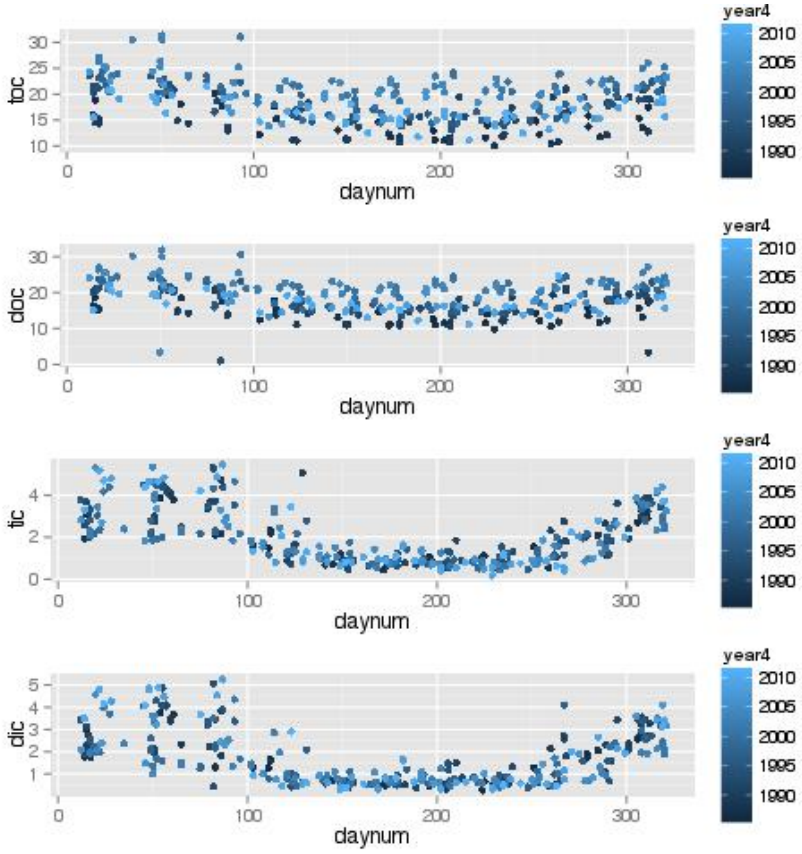

Figure S2-2: Dissolved and total carbon concentrations in Trout Bog from years 1986 through 2011. Concentrations were measured from surface samples. Axis and unit definitions: toc, total organic carbon [mg/l]; tic, total inorganic carbon [mg/l]; doc, dissolved organic carbon [mg/l]; dic, dissolved inorganic carbon [mg/l]; daynum, day of year.

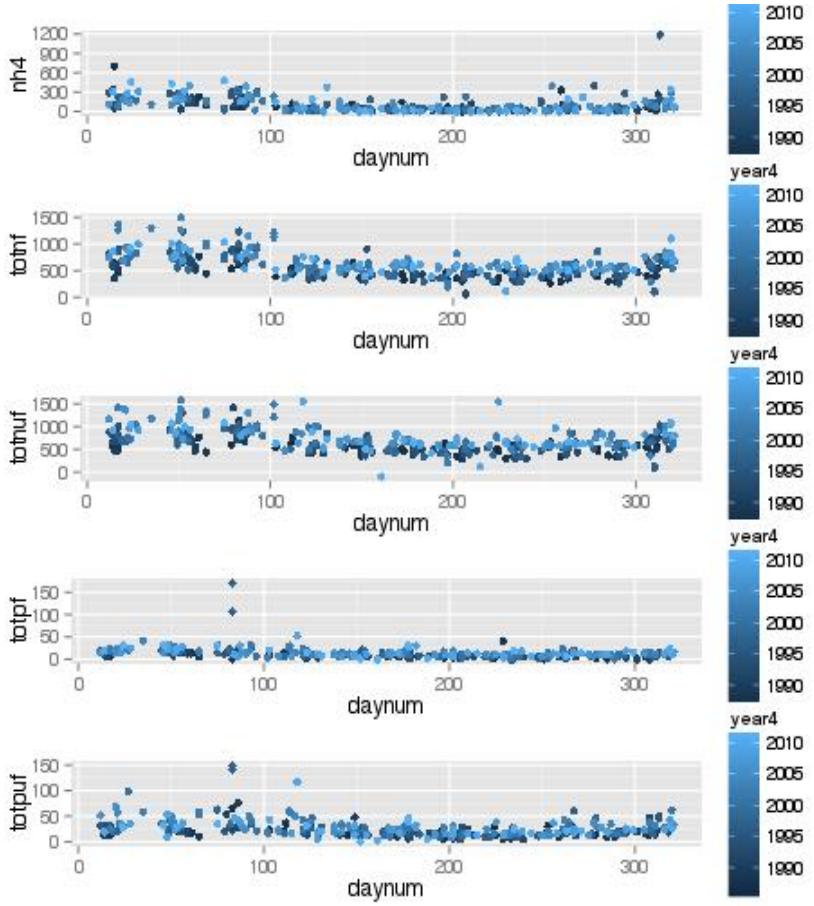

Figure S2-3: Dissolved and total nutrient concentrations in Trout Bog from years 1986 through 2011. Concentrations were measured from surface samples. Axis and unit definitions: nh4, ammonium [µg/l]; totnf, total nitrogen measured in samples filtered through 0.4 micron membrane [µg/l]; totnuf, total nitrogen measured in unfiltered sample [µg/l]; totpf, total phosphorous measured in samples filtered through 0.4 micron membrane [µg/l]; totpunf, total phosphorous measured in unfiltered samples [µg/l]; daynum, day of year.

Chapter 3: Metagenomic Inferences of Carbon Substrate Availability for Microbial Communities in Eutrophic and Humic Lakes

Authors:

Benjamin B. Crary, Leong-Keat Chan, Rex Malmstrom, Susannah Tringe, Stephanie Malfatti,

Tijana Glavina del Rio, Stefan Bertilsson, Katherine D. McMahon

Prologue

Chapter 3 is a draft manuscript that will be submitted for publications. The target journal is ISMEJ. Author contributions are as follows:

Benjamin Crary: Wrote original draft of the manuscript. Sequence analysis and statistical analyses.

Leong-Keat Chan: Assisted in sequence analysis.

Rex Malmstrom: Co-Principle Investigator.

Susannah Tringe: Co-Principle Investigator.

Stephanie Malfatti: Read assembly and annotation.

Tijana del Rio: Read Library Construction.

Stefan Bertilsson: Co-Principle Investigator.

Katherine D. McMahon: Principle Investigator.

Abstract:

Bacterial communities play an essential role in biogeochemical cycling in aquatic environments. The specific functions of discrete populations, however, are a much greater unknown. Aquatic bacterial genomes can be extremely streamlined and individual lineages seem to specialize in their carbon substrate use. Resource availability can differ dramatically among aquatic ecosystems, and here we aim to understand how the quality of dissolved organic carbon (DOC) in a eutrophic and a dystrophic lake can act as a driving force for bacterial community genomic content. In the dystrophic Trout Bog, we observed a higher number of polymer biodegrading enzymes (PBEs), indicating that there was a more abundant source of complex DOC and higher need for enzymes that can breakdown polymeric substances. There was also evidence for enrichment in amino acid transport machinery and aromatic degradation enzymes in Trout Bog, suggesting that the humic content is providing a labile source of substrates through photodegradation. In the eutrophic Lake Mendota, there were fewer overall DOC transporters per genome, but there were a higher proportion of sulfatase enzymes per PBEs. This higher proportion is consistent with the idea that the polymer sources in Lake Mendota are primarily of algal origin, and that the PBEs in this lake are targeting autochthonous carbon rather than allochthonous carbon.

Introduction

Bacterial communities assume a critical role in the many biogeochemical cycles that occur in aquatic ecosystems (Cotner and Biddanda 2002). Lakes are hot spots for remineralization of terrestrially and aquatically derived organic matter (OM), and have only recently been recognized as playing a significant role in the global carbon cycle (Cole et al. 2007). The relationship between OM and microbial community composition and function in freshwater lakes is therefore important for understanding carbon cycling on a global scale. Many geographical and environmental factors influence the quality of particulate and dissolved OM that is available to aquatic microbial communities. Such factors include the size, slope, and cover type of the surrounding watershed (Kratz et al. 1997) and trophic status (Toolan et al. 1991). Not surprisingly, heterotrophic bacterial communities appear to be structured by the relative contribution of allochthonous to autochthonous dissolved organic carbon (DOC)(Jones et al. 2009; Crump et al. 2003; Judd et al. 2006). Further, heterotrophic bacterial taxa appear to have preference to specific phytoplankton taxa, likely due to species-specific exudations of primary-produced OM (Kent et al. 2006).

Since the conception of the microbial loop model (Azam et al. 1983), slow but steady progress has been made in regards to understanding the role of specific aquatic bacterial taxa in the remineralization of organic compounds to inorganic forms. Much has been learned about which bacterial groups are ubiquitous and abundant in freshwater ecosystems using 16S rRNA gene-based methods (Newton et al. 2011). These include the acI lineage of the Actinobacteria phylum, the LD12 tribe of the Alphaproteobacteria, *Polynucleobacter* sp. of the Betaproteobacteria class, and the bac lineage of the phylum Bacteroidetes. Experimental work has begun to identify the specific organic substrates that these groups are capable of metabolizing. For example, many abundant aquatic bacteria have been observed to participate in the uptake of the chitin hydrolysis product, N-acetylglucosamine (Eckert et al. 2012; Nedoma et al. 1994). Subclusters of the *Polynucleobacter* genus have been grown on photodegradation products of DOM, including humic substances (Watanabe et al. 2009). Using FISH-MAR, studies have

determined that LD12 has a strong substrate preference for glutamate, glutamine, and glycine relative to the rest of the microbial assemblage; Actinobacteria and Betaproteobacteria were generally more active in amino acid uptake compared to *Flavobacteria* (Bacteroidetes); and independent clades or tribes of both *Polynucleobacter* and the acI Actinobacteria lineage have specific substrate uptake patterns (Salcher et al. 2011; Salcher et al. 2010; Buck et al. 2009; Hahn et al. 2012; Salcher et al. 2013). Thus, the activity and abundances of particular species may be driven in part by the availability of preferred substrates (Jones et al. 2009).

More recently, metagenomic studies have begun to highlight the functional and phylogenetic characteristics of freshwater ecosystems. Oh and colleagues observed a more copiotrophic metabolic strategy in lake samples compared to ocean samples (Oh et al. 2011). Metagenomic samples taken from Lac du Bourget linked carbohydrate metabolism to Bacteroidetes, nucleotide metabolism to Actinobacteria, and xenobiotic degradation to Alphaproteobacteria (Debroas et al. 2009), and Proteobacteria and nitrogen metabolisms were enriched in metagenomic samples taken during blooms of the toxic *Microcystis* within several lakes (Steffen et al. 2012).

Additionally, fungi have begun to be identified as potential participants in carbon cycling within pelagic ecosystems. Terrestrially evolved fungi are occasionally found in pelagic regions, presumably transported on terrestrial plant matter, but some fungi are indigenous to open water. Fungi have also been hypothesized to degrade otherwise inedible plant and algal detritus into labile forms for bacteria, subsequently contributing to the aquatic microbial loop (Gleason et al. 2008; Pabst et al. 2008).

In this study, we compared metagenomic datasets from two lakes with differing trophic status to further elucidate the role of bacterial and fungal taxa in freshwater carbon cycling. We predicted that 1) in Lake Mendota, a medium sized eutrophic lake, we would observe a larger abundance of genes involved in the metabolism of autochthonous carbon, 2) in Trout Bog, a small dystrophic lake, we would observe a larger abundance of genes involved in the metabolism of allochthonous carbon. We also explored the

hypothesis that fungi play a considerable role in the cycling of both autochthonous and allochthonous carbon. Glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs), and peptidases are all polymer biodegradation enzymes (PBEs), hence the abundance of genes encoding for these enzymatic groups was compared across the datasets to evaluate each microbial community's potential to metabolize allochthonous and autochthonous detritus. Amino acid and polyamine transporters were used to evaluate the metabolism of autochthonous carbon since these are expected to be abundant in algal exudates and the product of "sloppy" feeding by zooplankton and nanoflagelletes. Sulfatases were also considered as a marker of autochthonous polysaccharide utilization, as sulfated polysaccharides are a characteristic difference between aquatic and terrestrial eukaryotic cellular walls (McCandless and Craigie 1979). Carboxylic acid transporters and benzoyl-CoA were used as an indicator of humic substance degradation, as photodegradation products consist largely of carboxylic and aromatic structures (Bertilsson and Tranvik 1998). Taxonomic marker and single copy genes were used to identify and evaluate the abundance of bacteria and fungi in each ecosystem.

We also seek to contribute to the growing characterization of functional traits belonging to specific freshwater bacterial taxa, while proposing a basis for future investigation of fungal-bacterial cooperation in aquatic carbon cycling. There is currently a strong interest in carbon metabolism in comparative aquatic metagenomic studies, but few have evaluated the role of fungi. By presenting these data side by side, we point to the necessity of incorporating both fungal and bacterial analyses in future studies of aquatic carbon cycling in order to achieve a comprehensive understanding of the functional traits that are essential to carbon remineralization.

Methods:

Site Characteristics

Lake Mendota (ME) is a medium sized eutrophic lake located in southern Wisconsin (43.099, -89.405). This lake is fed by the Yahara Watershed, which is influenced heavily by agriculture and an urban environment, and experiences routine phytoplankton and cyanobacteria blooms. Trout Bog (TB) is a small dystrophic lake located in northern Wisconsin (46.041, -89.686). TB is a small humic seepage lake with a surrounding landscape that is dominated by boreal forests and a sphagnum mat. For physical and chemical lake characteristics see Supplemental Table S3-1.

Sampling Procedure

Five integrated epilimnetic samples were collected from the location of max depth in ME and TB. Three samples were collected from ME in 2010 (April 20, June 15, and October 29). Two samples were collected from TB in 2010 (June 3 and August 18). The calendar days were chosen to span across the ice-off season. All samples were taken at approximately 12:00 PM local time.

Integrated samples were collected from ME with vinyl tubing to a depth of 12 meters, as described by (Jones et al. 2007). For TB samples, temperature profiles taken with a handheld YSI A550 meter were used to define the thermal structure of the water column. Integrated epilimnetic samples were subsequently taken to the bottom of the thermocline with PVC tubes based upon the on site measurements, as described previously (Kent et al. 2004). Samples were kept in the dark and on ice for no more than 2 hours before they were filtered through 0.22 um membrane filters (Supor®-200, Pall Corporation) by vacuum filtration in the laboratory for biomass collection. Filters were then frozen and stored in a -80 °C freezer until further analysis. Chemical analysis was performed per the Northern Temperate Lakes Long Term Ecological Research program (http://lter.limnology.wisc.edu). Complete methodology can be found at http://lterquery.limnology.wisc.edu.

DNA Extractions and Sequencing

DNA extractions were performed using the FastDNA Spin Kit (FastDNA® SPIN KIT, MP Biomedicals) with minor adjustments. Briefly, a 10-minute incubation at room temperature with an addition of (25 μL x 10 μg/ml) lysozyme was performed prior to bead beating. After bead beating, a 60-minute incubation at 55 degrees C with an addition of (50 μL x 10 μg/ml) protease was performed.

Samples were sequenced on the Illumina HiSeq 2500 platform at the Joint Genome Institute in Walnut Creek, CA. Approximately 1 μg of DNA from each sample was used for sequencing. The DNA samples were amplified with ten rounds of PCR prior to library construction.

Assembly and Functional Annotations

Raw sequencing reads were quality filtered prior to merging paired ends and assembly. Low quality reads that had 80% or more of bases with a quality value of less than 20 were filtered out. Adapter sequences were also removed, and the remaining reads were merged with Fast Length Adjustment of Short Reads (FLASH) (Magoc and Salzberg 2011). In FLASH, the following criteria were used: mismatch value ≤0.25 and overlap bases ≥10. . Assembly of merged reads into contigs was performed by team at the Los Alamos National Laboratory (Los Alamos, NM). Contigs were deposited in the Integrated Microbial Genomes and Metagenomes (Expert Review) (IMG-MER) (Markowitz et al. 2012) under unique Taxon Object IDs (Table 3-1).

Cluster of Orthologous Groups (COG) annotations were performed on assembled reads (contigs) using RSP-BLAST on Position Specific Scoring Matrices provided by the Conserved Domain Database as part of the IMG standard analysis pipeline. Enzyme Commission (EC) and Protein Family (pfam) annotations were made analogously. Data was retrieved from the JGI's IMG-MER (Chen et al. 2013). Summary statistics are listed in Table 3-1.

Community Composition Assessment

Unassembled metagenomic reads encoding for the 16S rRNA gene were recruited with a database of nearly full-length 16S rRNA gene sequences from common freshwater lineages (Newton et al. 2011) using BLASTn with Blast+ (Altschup et al. 1990). High scoring pairs (HSPs) were only considered if it constituted at least a 70% match across a minimum length of 100 bp. Any such HSP was kept regardless of the recovered region within the 16S rRNA gene to avoid biases that arise from certain variable regions (Peura et al. 2012).

The 16S rRNA gene fragments that were recruited from each lake were pooled into single composite datasets for ME and TB. The composite 16S fragment samples were classified first against a custom curated database of common freshwater taxa using the Bayesian classifier in Mothur (Newton et al. 2011; Schloss et al. 2009). Reads that were not classified to the freshwater database with a confidence of 70 at the fifth taxonomic level were reclassified with the Greengenes database and the assignments were kept if the confidence was greater than 60 (DeSantis et al. 2006)

To estimate the average genome size, the methods descried by Frank & Sørensen and Raes et al. were adapted slightly (Frank and Sørensen 2011; Raes et al. 2007). Sequences corresponding to five single copy genes (rplA, rplC, rplD, rpsG, and rpsQ) were extracted from the STRING database, and unassembled metagenomic reads were recruited against this set using BLASTx. HSPs were kept if the E-value was less than 10E-3 and the alignment yielded at least a 50% match across a minimum length of 30 amino acid residues. To estimate the number of fungal genomes within the metagenomic datasets, each metagenome was compared against the Fungal Phylogenomic Database (Funybase) (Marthey et al. 2008). BLASTx was used to align a single HSP for each read and each HSP was regarded as a match if the bit score was greater than 60 and the % match was less than 50. The number of fungal genomes in the dataset was represented qualitatively by the number of hits to Funybase per mbp.

The 18S gene fragments were recruited from the metagenomes using the method described elsewhere (Tremblay and Tringe, in prep). Briefly, rRNA reads were recruited from the metagenomes using kmer signatures, and classified with the 18S references from Greengenes and Silva (Pruesse et al. 2007). Reads were classified to the deepest taxonomic level having a bootstrap confidence of 50%.

Pairwise Sample Comparisons

Each ME and TB metagenomic sample was compared pairwise using the Bray-Curtis dissimilarity metric based on the COG functional annotations (using relative abundance of each COG). Xipe-Totec was used to find statistical differences between samples (Supplemental Methods). The results from these analyses were interpreted to justify the use of composite lake samples for additional analysis.

Comparative Strategy

Comparisons between the TB and ME composite metagenomes were performed with three approaches. First, the abundances of PBEs and DOC transporters were compared as overall groups between the composite metagenomes. PBEs were quantified using EC annotations of GHs, CEs, and PLs (Cantarel et al. 2009), along with peptidases identified with EC annotations (Bairoch 2000). ABC-type DOC transporters were defined using COG annotations (Poretsky et al. 2010). Both PBE and DOC transporter abundances were reported relative to the average genome size in the composite metagenomes. Second, the COG annotations that were present in one lake but absent in the other were assessed for indicators of C substrate availability. Third, all COGs were ranked on abundance, and functions that were "overrepresented" (see below) were assessed for insights of C substrate availability.

Determining Criteria for COG overrepresentation

As noted above, datasets from each sample were pooled to create a single composite dataset for each lake. The COG assignments on each contig were adjusted for the contig average read depth to acquire an estimate of copy number within the original unassembled datasets. The estimated copies of each COG function were summed within each lake, and divided by the total number of assembled base pairs from the respective source lake. The summed and adjusted copy estimates were normalized by Z-score within each lake by taking the average and standard deviation of COG function copy estimates within the composite samples. Only COG functions that were identified in either ME or TB were used in the Z-scoring.

Differences in the Z-ranks of COGs in ME and TB were used to establish criteria for "overrepresentation". These criteria were designed to be loose enough to select a large enough subset of COGs such that there was adequate evidence discern differences across the ecosystems, while ensuring that the COGs were abundant and on long enough contigs to provide a phylogenetic estimate. The criteria were also constrained by computational considerations, in that the "overrepresented" subset had to be small enough to be manageable with the computational services available. With these guidelines, the following criteria were established.

COG Z-scores were first plotted on an xy-scatter plot. Next, the vertex of the scatter plot was identified at approximately (-0.6, -0.6). Lines with slopes of 3 and 1/3 (relating to 3X more abundant on either axis) were drawn from the vertex on the graph to establish lower bounds of abundance. Points that fell within these lines were excluded. The "Z-distance" for each COG between the two lakes was calculated by taking the absolute value of the difference of the two Z-scores. Those COGs with Z-distances that were less than one were excluded from further analysis. COGs that met these criteria were then considered "overrepresented" in one lake versus another.

Two COGs (COG1216 and COG0463) were identified as significantly overrepresented in TB using XIPE-Totec (Supplemental Methods) but did not meet our custom Z-scoring criteria. These two exceptions were therefore treated as if they did meet the Z-score criteria in further analyses.

Taxonomic Binning

The "Classifier for Metagenomic Sequences" known as ClaMs (Pati et al. 2011) was used for all taxonomic prediction based on contigs. While, in some cases, many contigs contained functional annotations that we were interested in binning (i.e. Overrepresented COGs, PBEs, DOC transporters), only contigs that were greater than 1 kbp could be accurately binned. Any contig less than 1 kbp was removed from phylogenetic analysis. In ClaMs, contigs were binned to the phylum level using the de Bruijn chain signature with a kmer length of 4. The actual distance cutoff was set to 0.01. The default taxonomy was used for binning at the phylum level, with the following minor adjustments; the largest contigs (>40kb) from 7 single cell amplified genomes of acI and 7 single cell amplified genomes of LD12 were included as reference bins. Additionally, the *Polynucleobacter* (Pnec) genome that is default in ClaMs was used as a bin, due to its known high abundance in TB. After a best phylogenetic bin was calculated for each contig, the results were weighted based on the estimated average read depth per each contig. Consider the following example; COGX is estimated to have 20 copies in TB, and this COG is on 10 contigs. If 1/10 contigs binned to Pnec, and that particular contig had an estimated copy number of 5, then 5 of the 20 copies of COGX in TB came from a Pnec genome.

Results:

16S rRNA and 18S rRNA-based Taxonomic Distribution

In order to gain a broad view of community composition, we extracted and taxonomically identified metagenomic reads that clearly mapped to 16S rRNA genes ("16S-reads" henceforth), using a custom curated training set focused on freshwater lineages. More than 97% of the 16,445 bacterial 16S rRNA gene fragments could be assigned to the phylum level (Figure3-1A). Proteobacteria was the most abundant phylum in both lake datasets, making up 33% and 38% of the 16S-reads, respectively. The abundance of reads affiliated with Alphaproteobacteria and Betaproteobacteria were very similar in ME (14.7% and 14.3% of total reads, respectively), but Betaproteobacteria were considerably more abundant than Alphaproteobacteria in reads from TB (23% and 8% of the 16S-reads, respectively).

In ME, the most abundant tribe was LD12 (Alphaproteobacteria) (Figure 3-1B). LD12 made up more than 8% of the 16S-reads in ME, but less than 1% in TB. A species of *Acetobacteracaea* was the most abundant Alphaproteobacteria in TB (Supplemental Table S3-2). Alternatively, *Polynucleobacter* subcluster C (*PnecC*) (Betaproteobacteria) was the largest constituent of the TB community, contributing more than 8%, but less than 1% of the 16S-reads in ME. Another Betaproteobacteria clade, betI-A, made up a relatively abundant 4.5% and 3.0% of the ME and TB communities, respectively. Tribes within this clade (Lhab A-1, Lhab A-2, Lhab A-3, and Lhab A-4) and the other Proteobacteria clades (betIII, betIV, and alfIV) were observed in both lakes with few discernable patterns of abundance.

The ME metagenomes were made up of a larger percentage of Actinobacteria than the metagenomes from TB (28% and 20%, respectively), and there was distinct tribe-level partitioning within this phylum. Two tribes in particular, acI-B1 and acI-B2, appear to have a strong preference for ME and TB, respectively. The tribes acV-A2 and acI-A6 were also observed to have strong partitioning to TB and ME, respectively.

Reads affiliated with Cyanobacteria comprised 6.9% of the ME 16S-reads, while only 0.9% of the TB community. In ME, the dominant class was Nostocophycideae, but assignments beyond this class were largely unclassified. Previous studies show the cyanobacteria community within ME to be largely comprised of *Aphanizominon* and *Microcystis*, which frequent population succession (Beversdorf, Miller, and McMahon 2013).

Two taxonomic groups within the phylum Chlorobi were abundant in TB (4.1% and 4.0%, respectively) but rare in ME: the genus *Pelodictyon* and another unclassified lineage of the class Chlorobiaceae. *Pelodictyon* is a green sulfur bacterium often found in stratified waters where light can penetrate to sulfide containing regions of the water column (Savvichev et al. 2005; Pfennig and Cohen-Bazire 1967).

Plancotmycetes and Acidobacteria each contributed <5% to each lake's 16S-reads, but there was strong partitioning between the lakes. Acidobacteria-affiliated reads comprised 4.7% of the TB 16S-reads while <0.5% of the ME 16S-reads. The family Holophagaceae made up 94% of the Acidobacteria assignments. Planctomycetes were detected at 2.0% of the ME 16S-reads, but were not observed in the TB 16S-reads. Planctomycetia was the most abundant class within Planctomycetes, but no single group dominated the abundance of that class.

Fragments of the 18S rRNA gene were also considered for taxonomic placement (Figure 3-1C). The most abundant fungal genus, in both lakes, was the genus *Sebacinales* (Agarimycotina), which was recently described as a ubiquitous endophyte without any coherent ecological occurrence patterns (Weiß et al. 2011). Additional abundant fungal groups included uncharacterized members from the class Ascomycota and the genus *Agaricostilbales*.

Overrepresented COGs

A Bray-Curtis pairwise comparison between functional content of all samples revealed that there was more differences across space than time (Figure 3-2B). Previous studies have also demonstrated greater spatial differences across lakes than temporal differences within (Yannarell et al. 2003). Thus pooling of lake samples was deemed appropriate to identify the largest differences between ecosystems, independent of time. From this point forward, the mention of TB and ME refers to the composite metagenomic samples (unless stated otherwise).

Using the custom criteria we established for "overrepresentation", we identified 86 COGs that were overrepresented in TB, while only 14 in ME (Figure 3-2A). Of the COGs that were overrepresented in TB, a large fraction (37/86) were of general function prediction only or an unknown function, categorically (Table 3-2). Some nonspecific functional predictions could be made for these COGs, though few elucidated much information on substrate preference or other traits that might define TB communities.

Though many overrepresented genes were unidentifiable in TB, perhaps owing to the fact that bog lakes are traditionally less studied than eutrophic lakes, a few COG annotations revealed differences in substrate availability across lakes. Several COGs overrepresented TB, notably a lysozyme (COG3772), a chitinase (COG3179), and a putative secretion enzyme with lytic characteristics (COG3926) (Pei and Grishin 2005), were all indicators of cell wall degradation. COG0677 may also indicate the use of mannose and other polysaccharides as a C acquisition strategy in TB. In ME, a polyamine transporter (COG0687) and an amino acid derivative dehydrogenase (COG1748) were overrepresented, along with a glucose dehydrogenase (COG2133).

Phylogenetic Binning of Overrepresented COGs

A comparison of the phylogenetic distribution of overrepresented COGs in each lake revealed that bacteria belonging to the phyla Bacteroidetes, Cyanobacteria, and Firmicutes contributed a large portion of the COGs that were overrepresented in both TB and ME (figures 3-3A and 3-3B). However, it is important to note that distinct clades likely contributed to these COGs in each lake. For example, the Bacteroidetes bacI-B clade was abundant in TB, while bacII-A and bacI-A were abundant in ME (Figure 3-1). This is consistent with the notion that only a few phyla are commonly found in freshwater lakes, but that ecologically distinct clades within the phyla inhabit different niches and therefore different lakes.

While these three phyla appeared to fill much of the unique niche space, as evidenced by their contributions to overrepresented genes, Chlorobi and the Actinobacteria acI lineage appeared to also harbor genetic content that was more abundant in TB and ME, respectively. The acI Actinobacteria contributed much of the overrepresented genetic content in ME, of which COG0687 and COG0538 (amino acid metabolism and energy production, respectively) are perhaps most notable. Alternatively, the Chlorobi appear to contribute a specific set of the overrepresented COGs in TB including three COGs comprising an ABC-type nitrate/sulfonate/bicarbonate transport system (COG0600, COG0715, and COG1116) and a nitrogenase protein (COG2710).

Polynucleobacter (Pnec) appears to share the same ABC-type transport system with the Chlorobi, as a substantial proportion of the contigs containing these COGs binned to this clade. Contigs containing a predicted chitinase (COG3179) were clearly binned as Pnec, along with contributions from Cyanobacteria, LD12, and other phyla. Additionally, a phage related lysozyme (COG3772) binned strongly to Pnec.

Analysis of Functional Presence/Absence

A presence/absence comparison between the two sample sets revealed that 11.0% of the 3650 COGs that were identified in ME were not detected at all in TB (Figure 3-2A, Supplemental Table S3-3). Similarly 11.0% of the COGs that were identified in TB were not identified in ME. The COG functions that were observed in ME but not TB included several components of a nitrate reductase (COG1140, COG2180) and a lignostilbene related enzyme (COG3670). Of those COG functions that were not detected in ME, but which were observed in TB, were a benzoyl-CoA reductase (COG1775), and a uncharacterized distant relative of cell wall hydrolases (COG3863), though many differentially represented functions were uncharacterized. One of six contigs containing COG1175 annotation was long enough (>1 kb) to provide a taxonomic estimate with ClaMS, and the contig was most closely affiliated with a Chlorobi. COG3863 was considerably abundant (Z-score=1.29) in Trout Bog, and the functional annotation for this gene came to a protein homolog match from an Actinomycetales. On many contigs, COG3863 preceded a gene annotated as an N-acetylmuramoyl-L-alanine amidase, or a member of the cysteine/histidine/amidohydrolase protein (CHAP) domain.

PBEs and ABC-Type DOC Transporters

A survey of PBEs revealed an elevated number of these genes per genome within TB (Table 3-3). CEs and GHs were the most distinguishable between TB and ME. These two enzymatic functions had over 94% more copies per genome in TB, compared to ME. Additionally, PLs were more represented in ME, though there were less than 0.1 copies/genome in either lake. Peptidases were 20% more abundant in TB, while sulfatases were observed to have equal copies per genome in each lake.

By binning contigs that contained these GHs, several taxonomic groups (i.e.: Fungi, Bacteroidetes, Firmicutes, Cyanobacteria) appeared to be most enzymatically capable of degrading

complex molecules in both lakes, yet Fungi and Chlorobi appear to be much greater participants in TB than in ME (Figure 3-4). Sulfatases binned to the same groups in each lake, with the exception of very few sulfatases binning to Chlorobi in TB.

A comparison of the ABC-type DOC transporters showed that both lakes have similar overall strategies for DOC transport (Table 3-3). The largest differences between the two lakes were the higher number of amino acid transporters per genome in TB and a higher number of polyamine transporter copies per genome in ME. Compatible solute transporters were more abundant in TB, but the remaining transporter types were in slightly higher abundance within ME.

Discussion:

Organic Acid Transport and Metabolism

In this study we aimed to identify the unique functional characteristics and C substrate availability of lakes with varying trophic statuses. We investigated the largest functional differences across the ecosystems for insights into substrate preferences, and we targeted PBEs and DOC transporters for further characterization of the two lakes. Microbial communities are the primary decomposers of organic matter in aquatic ecosystems, and thus understanding how lake trophic status can influence the metabolic function of a microbial community will enable more accurate assessment of the net C metabolism of a lake.

The eutrophication of ME and its effect on increased primary production has been well researched (Pedrós-Alió and Brock 1982; Soranno et al. 2013; Beversdorf et al. 2013). The overall mass of primary production that is taken up by heterotrophic bacteria is likely to be noteworthy in eutrophic lakes, though the fraction may appear minor in eutrophic lakes (~5%) compared to oligotrophic lakes (~40%) (Lignell 1990). In addition, heterotrophic bacteria benefit from the release of organic matter

during zooplankton feeding (Jumars et al. 1989). Phytoplankton communities release exudates during their life cycle, and although specific exudate quality varies between phytoplankton species, the composition of exudates is typically acidic and sulfated polysaccharides, amino acids and polyamines (Bahulikar and Kroth 2008; Grossart and Simon 2007).

Contrary to our prediction that amino acid transport would be more important in ME because of increased algal activity and their associated exudates, we observed that genes targeting this portion of the DOC pool were more prevalent in TB (Table 3-3). Owing to the fact that DOC in humic lakes is highly complex and poorly characterized, the ability to effectively acquire labile DOC (e.g. amino acids, carboxylic acids and carbohydrates) as it becomes available would provide an advantage to any such population given that the supply of that labile DOC warrants that ecological strategy. Amino acids within TB could be derived from phytoplankton, but it now appears more probable that amino acids are becoming freed from humic material during photodegradation in a high enough abundance and regularity to make them an effective growth substrate in TB (Jörgensen et al. 1998).

Given the overall higher abundance of these transporters, it was surprising that none of the COGs we identified as "overrepresented" in TB were amino acid transporters (Figure 3-2A, Table 3-2). This suggests a generalist amino acid usage strategy and an irregular pattern to the production of specific amino acid substrates in this ecosystem. An added advantage of this generalist strategy would be the acquisition of labile nitrogen. In TB, it is inferred that the high nutrient content is deceiving, and that most of the nutrients are bound to complex humic materials (Sachse et al. 2001; Burkert et al. 2003). One way of efficiently acquiring these nutrients at the same time as acquiring carbon would be to target smaller photodegradation products such as amino acids.

Overall, photodegradation products of humic material (largely fatty acids, organic acids, and alcohols) may be a source that could provide enough for the entire bacterial community (Bertilsson and Tranvik 1998; Jörgensen et al. 1998; Tarr et al. 2001). PnecC (>8% of TB community) is one strain of

bacteria that may be using this source of organic carbon. Watanabe et al. observed Pnec growth in isolation when fed photochemically degraded organic carbon, specifically organic acids, and Hahn et al. have characterized assimilation patterns (Watanabe et al. 2009). Binning of DOC transporters revealed that Pnec is highly enriched in carboxylic transporters along with amino acid transporters, to a lesser degree. Interestingly, the overrepresented COG that binned most strongly to PnecC was an uncharacterized gene in TB and was the fourth most abundant COG (COG3181) within the TB dataset.

The vast majority of gene copies annotated as COG3181 were predicted as the periplasmic component (tctC) of the tripartite tricarboxylate transporter (TTT) family. tctC is specific to fluorocitrate, citrate, and isocitrate with micromolar affinities, while cation inhibitors and dependencies are poorly defined (Winnen et al. 2003). The inability of Pnec strain F10 to consume citrate (a photooxidation product of humic substances) corroborates previous assertions of Pnec's intragenus diversity (Hahn, Scheuerl, et al. 2012), and the overrepresentation of tctC in TB could indicate an ecological strategy conserved in this ecosystem.

The detection of COG1775 (Benzoyl-CoA reductase) in TB, and its absence in ME lends additional credence to humic matter being utilized. This enzyme functions as a degradation pathway of aromatic compounds, which are commonly found in humic acids. COG1775 was found on six different contigs, and Chlorobi was the best match for the single contig that was long enough to bin with ClaMs.

Polyamine Transport and Metabolism

While enzymes involved in transporting and metabolizing photodegradation products of humic substances (including amino acids) appeared to be highly enriched within TB, Polyamine transporters were observed in higher abundances in ME, and a transporter protein specific to spermidine/putrescine (COG0687) was overrepresented in ME based on our strict criteria (Figure 3-2, Table 3-2). In addition,

there was a saccharopine dehydrogenase (COG1748) that was overrepresented in ME. These two COGS both binned strongly to acI, one of the most abundant clades (Figure 3-1B), suggesting a niche for specialists based on particular polyamines or amino acid substrates that is absent in TB. Based on the presence of polyamine chains in diatom cell walls, and the history or diatom blooms occurring ME, it is possible to speculate that the acI clade is adapted to acquire diatom detritus in ME (Kröger 1999; Meloche et al. 1938).

While recent studies using single cell genomes of acI-B1 show that putrescine/spermidine uptake may be a core trait in this lineage, this metagenomic evidence further illustrates that acI is likely to specialize in metabolizing autochthonous amino acids and polyamines ((Garcia et al. 2013), Ghylin, in prep). Though still unclear in the literature, it is interesting to speculate whether the acI-B2 tribe also contains this COG function. If acI-B2 did have this gene, then we would not have observed it to be overrepresented in ME given the similar abundances of acI-B1 in ME and acI-B2 in TB. Previous studies identified pH as strongly correlated to the abundance of these two tribes (Newton et al. 2007), but our results suggest that C substrate preference may also be a driver. Other hallmarks of potential major advantages within the acI genome should be explored if acI-B2 indeed lacks this capability, in order to explain acI-B abundance and persistence in both dystrophic and eutrophic lakes.

Polymer Degradation

The survey performed on PBEs (GHs, PLs, CEs, peptidases, and sulfatases) mostly aligned with our expectations (Table 3-3, Figure 3-4). Peptidases were equally abundant in both lakes, but there were more copies of GHs, PSs, and CEs per genome in TB, validating the assumption that the TB ecosystem selects for community components that are capable of degrading polymeric compounds. The relative lack of sulfatases among genomes in TB (lower sulfatase copies per genome compared to other PBEs) further indicated that the source of the polymeric substances is not autochthonous, as algal cell walls tend to

contain higher amounts of sulfated compounds (Glöckner et al. 2003). In ME, where we expected the complex molecule degrading enzymes to be targeting phytoplankton derived polymers, there was in fact a higher conservation of sulfatases in relation to PBEs.

Several COGs that were overrepresented in TB indicated that this community could effectively target polysaccharides as a substrate sources. Particularly, there were two lytic enzymes (COG3772 and COG3926), as well as a chitinase (COG3179) overrepresented. Chitin, found in cell walls of eukaryotes, may be an indication of a bacterial community that utilizes cell wall components of phytoplankton, fungal, or leaf detritus. The overrepresentation of COG0677, which is putatively active in the degradation of a storage polysaccharide, indicates that cell wall components are not the sole source of polysaccharides in this environment. While there were fewer relative sulfatases in TB, there is certainly evidence demonstrating that bacteria in bog lakes are largely influence by phytoplankton regimes (Kent et al. 2006). Recent literature also suggests that algal growth stimulates heterotrophic activity and decomposition of leaf litter (Kuehn et al. 2013), thus it is possible that the autochthonous C production within TB is prompting the respiration of allochthonous polymers given the raw abundance of PBEs.

The taxonomic distribution of PBEs in TB and ME indicate that that general strategies are conserved within phyla across ecosystems, but specific clades are adapted to resource availability. Chlorobi, Bacteroidetes, and Firmicutes appear to be the most capable of general polymer degradation, yet Bacteroidetes and Firmicutes appeared to maintain a larger number of peptidases in ME. Bacteroidetes and Firmicutes seem to specialize in acquiring polyamines in ME, a strategy that was overall more abundant in this system. The Firmicutes did appear to maintain a considerably higher number of lipid transporters, despite the other noted functional similarities between Firmicutes and Bacteroidetes.

Bacteroidetes and Firmicutes also contribute much of the material that is overrepresented in either lake, illustrating the diversity of functional potential that different clades within the same phylum can

exhibit. Of course, the large genome size of members of these phyla may also skew the interpretation of their "unique" functions across ecosystems, because the overrepresented material may perform extraneous or auxiliary functions that streamlined genomes simply are not maintaining. Regardless of their levels of expression, however, these overrepresented genes must persist in the environment for some evolutionary and ecological purpose. It is for this reason that it is, perhaps, worth looking to Bacteroidetes, Firmicutes, and Cyanobacteria (and other organisms with large genomes) as indicators of some of the less obvious features of an ecosystem, and looking to the streamlined organisms as indicators of the fundamental features of an ecosystem.

Amplification of a diverse set of single cells revealed that "wild-type" have fewer gene duplicates and paralogs than cultured representatives in oligotrophic environments (Swan et al. 2013), yet there is still an approximate 3-fold difference between the smallest genomes (acI-B1 ~ 1.2 mbp (Garcia et al. 2013)) and the average estimated genome sizes (Table 3-1). Thus, there is a large range of genome streamline strategies remaining in these lakes, and this range is probably larger for eutrophic and dystrophic environments than oligotrophic environments (Raes et al. 2007). Subsequently, we can look at the genes that binned to the acI-B1 clade, which has a small genome and is highly abundant in ME but almost absent in TB (Figure 3-1B), and interpret that polyamine availability is a common ecosystem feature in eutrophic lakes but not dystrophic lakes. Alternatively we can see that Bacteroidetes, Firmicutes, and Cyanobacteria all maintain COG5323 in TB, which is overrepresented but have been poorly characterized in reference datasets. Presumably, we could garner information on more the more subtle functions present or absent in both lakes by investigating this COG with functional genomics strategies.

Aquatic Fungi

The contribution of fungi was perhaps most relatable to that of Bacteroidetes, Firmicutes, and Chlorobi in TB, and Bacteroidetes and Firmicutes in ME based on the taxonomic binning of both PBEs and DOC transporters (Figure 3-4). Each of these phylogenetic groups appeared to maintain a diverse set of PBEs, but Fungi and Chlorobi appeared to maintain fewer sulfatases relative to other PBEs than either Firmicutes or Bacteroidetes within both lakes (Figure 3-4). Recent findings by Kuehn et al. found that fungi degradation of leaf litter was stimulated by algal growth in aquatic ecosystems, so it should not be surprising that the functional potential of fungi most closely resembled two phyla (Firmicutes and Bacteroidetes) that are commonly associated with algal bloom and busts (Kuehn et al. 2013). Fungi also contributed to a variety of overrepresented functions in both ecosystems; however, this should not be surprising given the broad diversity of organisms that are classified as fungi.

The characteristics of bog lakes such as large depositions of terrestrial plant matter, low pH, and shallow depth, allow fungi to thrive in such a system, yet more single copy fungal genes were observed in ME (Table 3-1). By classifying the 18S gene fragments, we observed that the majority of fungi in TB and ME were of the genus *Sebacinales* (Figure 3-1). This genus is considered ubiquitous, yet most commonly an endophyte (Weiß et al. 2011). Given that there is a lack of reference 18S sequences for aquatic fungi (Begerow et al. 2010), we suspect that there could be improvement of Eukaryotic assignment upon the acquisition of near full-length 18S sequences. Ascomycota and Basidiomycota have been identified previously in bogs and fens (Lin et al. 2012), and extensive reviews indict chytrids as a present and active in aquatic ecosystems (Gleason et al. 2008).

This study further contributes to the growing understanding of the specific role of freshwater microbial communities in allochthonous and autochthonous carbon cycling. Here, many genes were identified as overrepresented in either eutrophic ME or dystrophic TB, yet most gene annotations yielded only broad functional predictions. The depth of metagenomic data presented here provided a launching

point to determine to which taxonomic groups these abundant yet uncharacterized genes belong, and we suggest that continued focus be given to the investigation of these common freshwater bacteria, while introducing an emphasis on aquatic bacterial-fungal interactions to elucidate the dynamics in carbon and nutrient cycling.

Acknowledgements

We thank staff and students at Trout Lake Station and the Center for Limnology at the University of Wisconsin Madison for logistical support for sample collection. We thank the Joint Genome Institute for supporting this work through the Community Sequencing Program, performing the bioinformatics, and providing technical support. We would also like to acknowledge the Biotechnology Training Program of the National Institutes of Health at the University of Wisconsin-Madison for providing financial support for TWG's research and training (grant #5T32GM08349). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. KDM acknowledges funding from the United States National Science Foundation Microbial Observatories program (MCB-0702395), the Long Term Ecological Research program (NTL-LTER DEB-0822700), and a CAREER award (CBET-0738309).

References

Altschup, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215: 403–410.

Azam, F., T. Fenchel, J. G. Field, J. S . Gray, L. A. Meyer-Reil, and F. Thingstad. 1983. "The Ecological Role of Water-Column Microbes in the Sea." *Marine Ecology* 10: 257–263.

Bahulikar, Rahul a., and Peter G. Kroth. 2008. "The Complex Extracellular Polysaccharides of Mainly Chain-Forming Freshwater Diatom Species From Epilithic Biofilms 1." *Journal of Phycology* 44 (6) (December): 1465–1475.

Bairoch, a. 2000. "The ENZYME Database in 2000." *Nucleic Acids Research* 28 (1) (January 1): 304–5.

Begerow, Dominik, Henrik Nilsson, Martin Unterseher, and Wolfgang Maier. 2010. "Current State and Perspectives of Fungal DNA Barcoding and Rapid Identification Procedures." *Applied Microbiology and Biotechnology* 87 (1) (June): 99–108.

Bertilsson, Stefan, and Lam J Tranvik. 1998. "Photochemically Produced Carboxylic Acids as Substrates for Freshwater Bacterioplankton" 43: 885–895.

Beversdorf, Lucas J, Todd R Miller, and Katherine D McMahon. 2013. "The Role of Nitrogen Fixation in Cyanobacterial Bloom Toxicity in a Temperate, Eutrophic Lake." *PloS One* 8 (2) (January): e56103.

Buck, Ulrike, Hans-Peter Grossart, Rudolf Amann, and Jakob Pernthaler. 2009. "Substrate Incorporation Patterns of Bacterioplankton Populations in Stratified and Mixed Waters of a Humic Lake." *Environmental Microbiology* 11 (7) (July): 1854–65.

Burkert, Ulrike, Falk Warnecke, Dieter Babenzien, Elke Zwirnmann, and Jakob Pernthaler. 2003. "Members of a Readily Enriched Beta-Proteobacterial Clade Are Common in Surface Waters of a Humic Lake." *Applied and Environmental Microbiology* 69 (11): 6550–6559.

Cantarel, Brandi L, Pedro M Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. 2009. "The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics." *Nucleic Acids Research* 37 (Database issue) (January): D233–8.

Chen, I-Min A., Victor M Markowitz, Ken Chu, Iain Anderson, Konstantinos Mavromatis, Nikos C Kyrpides, and Natalia N Ivanova. 2013. "Improving Microbial Genome Annotations in an Integrated Database Context." *PloS One* 8 (2) (January): e54859.

Cole, J. J., Y. T. Prairie, N. F. Caraco, W. H. McDowell, L. J. Tranvik, R. G. Striegl, C. M. Duarte, et al. 2007. "Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget." *Ecosystems* 10 (1) (February 13): 172–185.

Cotner, James B, and Bopaiah A Biddanda. 2002. "Small Players, Large Role: Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems." *Ecosystems* 5 (2): 105–121.

Crump, Byron C, George W Kling, Michele Bahr, and John E Hobbie. 2003. "Bacterioplankton Community Shifts in an Arctic Lake Correlate with Seasonal Changes in Organic Matter Source." *Applied and Environmental Microbiology* 69 (4): 2253–2268.

Debroas, Didier, Jean-François Humbert, François Enault, Gisèle Bronner, Michael Faubladier, and Emmanuel Cornillot. 2009. “Metagenomic Approach Studying the Taxonomic and Functional Diversity of the Bacterial Community in a Mesotrophic Lake (Lac Du Bourget--France).” *Environmental Microbiology* 11 (9) (September): 2412–24.

DeSantis, T Z, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. 2006. “Greengenes, a Chimera-checked 16S rRNA Gene Database and Workbench Compatible with ARB.” *Applied and Environmental Microbiology* 72 (7) (July): 5069–72.

Eckert, Ester M, Michaela M Salcher, Thomas Posch, Bettina Eugster, and Jakob Pernthaler. 2012. “Rapid Successions Affect Microbial N-acetyl-glucosamine Uptake Patterns During a Lacustrine Spring Phytoplankton Bloom.” *Environmental Microbiology* 14 (3) (March): 794–806.

Frank, Jeremy a, and Søren J Sørensen. 2011. “Quantitative Metagenomic Analyses Based on Average Genome Size Normalization.” *Applied and Environmental Microbiology* 77 (7) (April): 2513–21.

Garcia, Sarahi L, Katherine D McMahon, Manuel Martinez-Garcia, Abhishek Srivastava, Alexander Sczyrba, Ramunas Stepanauskas, Hans-Peter Grossart, Tanja Woyke, and Falk Warnecke. 2013. “Metabolic Potential of a Single Cell Belonging to One of the Most Abundant Lineages in Freshwater Bacterioplankton.” *The ISME Journal* 7 (1) (January): 137–47.

Gleason, Frank H, Maiko Kagami, Emilie Lefevre, and Telesphore Sime-ngando. 2008. “The Ecology of Chytrids in Aquatic Ecosystems : Roles in Food Web Dynamics.” *English* 22: 17–25.

Glöckner, F O, M Kube, M Bauer, H Teeling, T Lombardot, W Ludwig, D Gade, et al. 2003. “Complete Genome Sequence of the Marine Planctomycete Pirellula Sp. Strain 1.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (14) (July 8): 8298–303.

Grossart, Hp, and M Simon. 2007. “Interactions of Planktonic Algae and Bacteria: Effects on Algal Growth and Organic Matter Dynamics.” *Aquatic Microbial Ecology* 47 (3) (May 16): 163–176.

Hahn, Martin W, Arevik Minasyan, Elke Lang, and Ulrike Koll. 2012. “Polynucleobacter Difficilis Sp . Nov ., a Planktonic Freshwater Bacterium Affiliated with Subcluster B1 of the Genus Polynucleobacter.” *International Journal of Systematic and Evolutionary Microbiology*: 376–383.

Hahn, Martin W, Thomas Scheuerl, Jitka Jezberová, Ulrike Koll, Jan Jezbera, Karel Šimek, Claudia Vannini, Giulio Petroni, and Qinglong L Wu. 2012. “The Passive yet Successful Way of Planktonic Life: Genomic and Experimental Analysis of the Ecology of a Free-living Polynucleobacter Population.” *PloS One* 7 (3) (January): e32772.

Jones, Stuart E, Ryan J Newton, and Katherine D McMahon. 2009. “Evidence for Structuring of Bacterial Community Composition by Organic Carbon Source in Temperate Lakes.” *Environmental Microbiology* 11 (9) (September): 2463–72.

Jones, Stuart E, Ashley L Shade, Katherine D McMahon, and Angela D Kent. 2007. “Comparison of Primer Sets for Use in Automated Ribosomal Intergenic Spacer Analysis of Aquatic Bacterial Communities: An Ecological Perspective.” *Applied and Environmental Microbiology* 73 (2) (January): 659–62.

Judd, Kristin E., Byron C. Crump, and George W. Kling. 2006. "Variation in dissolved organic matter controls the bacterial poduction and ocmmunity composition" *Ecology* 87 (8): 2068–2079.

Jumars, Peter A, Deborah L Penry, John A Baross, Mary Jane Perry, and Bruce W Frost. 1989. "Closing the Microbial Loop : Dissolved Carbon Pathway to Heterotrophic Bacteria from Incomplete Ingestion , Digestion and Absorption in Animals." *Deep-Sea Research* 36 (4): 483–495.

Jörgensen, Niels O G, Lars Tranvik, Helene Edling, Wilhelm Grane, and Lindell Mans. 1998. "Effects of Sunlight on Occurrence and Bacterial Turnover of Specific Carbon and Nitrogen Compounds in Lake Water." *FEMS Microbiology Ecology* 25: 217–227.

Kent, A D, S E Jones, A C Yannarell, J M Graham, G H Lauster, T K Kratz, and E W Triplett. 2004. "Annual Patterns in Bacterioplankton Community Variability in a Humic Lake." *Microbial Ecology* 48 (4) (November): 550–60.

Kent, A D, S E Jones, G H Lauster, J M Graham, R J Newton, and K D McMahon. 2006. "Experimental Manipulations of Microbial Food Web Interactions in a Humic Lake: Shifting Biological Drivers of Bacterial Community Structure." *Environmental Microbiology* 8 (8) (August): 1448–59.

Kratz, Timothy K., Katherine E. Webster, Carl J. Bowser, John J. Magnuson, and Barbara J. Benson. 1997. "The Influence of Landscape Position on Lakes in Northern Wisconsin." *Freshwater Biology* 37: 209–217.

Kröger, N. 1999. "Polycationic Peptides from Diatom Biosilica That Direct Silica Nanosphere Formation." *Science* 286 (5442) (November 5): 1129–1132.

Kuehn, Kevin A, Steven N Francoeur, Robert H Findlay, and Robert K Neely. 2013. "Priming in the Microbial Landscape: Periphytic Algal Stimulation of Litter-associated Microbial Decomposers." *Ecology* (August 10). doi:10.1890/13-0430.1. http://dx.doi.org/10.1890/13-0430.1.

Lignell, Risto. 1990. "Exretion of Organic Carbon by Phytoplankton: Its Relation to Algal Biomass, Primary Productivity and Bacterial Secondary Productivity in the Baltic Sea." *Marine Ecology Progress Series* 68: 85–99.

Lin, X, S Green, M M Tfaily, O Prakash, K T Konstantinidis, J E Corbett, J P Chanton, W T Cooper, and J E Kostka. 2012. "Microbial Community Structure and Activity Linked to Contrasting Biogeochemical Gradients in Bog and Fen Environments of the Glacial Lake Agassiz Peatland." *Applied and Environmental Microbiology* 78 (19) (October): 7023–31.

Magoc, Tanja, and Steven L Salzberg. 2011. "FLASH : Fast Length Adjustment of Short Reads to Improve Genome Assemblies." *Bioinformatics* 27 (21): 2957 –63.

Markowitz, Victor M, I-Min a Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Yuri Grechkin, Anna Ratner, et al. 2012. "IMG/M: The Integrated Metagenome Data Management and Comparative Analysis System." *Nucleic Acids Research* 40 (Database issue) (January): D123–9.

Marthey, Sylvain, Gabriela Aguileta, François Rodolphe, Annie Gendrault, Tatiana Giraud, Elisabeth Fournier, Manuela Lopez-Villavicencio, Angélique Gautier, Marc-Henri Lebrun, and Hélène Chiapello. 2008. "FUNYBASE: a FUNgal phYlogenomic dataBASE." *BMC Bioinformatics* 9 (i) (January): 456.

McCandless, E L, and J S Craigie. 1979. "Sulfated Polysaccharides in Red and Brown Algae." *Annual Review of Plant Physiology* 30 (1): 41–53.

Meloche, V. W., G. Leader, L. Safransk, and C. Juday. 1938. "The Silica and Diatom Contenet of Lake Mendota Water." *Wisconsin Academy of Sciences, Arts, and Letters* 31: 363–376.

Nedoma, J., J. Vrba, J. Hejzlar, K. Simek, and V. Straskrabova. 1994. "N-acetylglucosamine Dynamics in Freshwater Environments: Concentration of Amino Sugars, Extracellular Enzyme Activities, and Microbial Uptake." *Limnology and Oceanography* 39 (5): 1088–1100.

Newton, Ryan J, Stuart E Jones, Alexander Eiler, Katherine D McMahon, and Stefan Bertilsson. 2011. "A Guide to the Natural History of Freshwater Lake Bacteria." *Microbiology and Molecular Biology Reviews : MMBR* 75 (1) (March): 14–49.

Newton, Ryan J, Stuart E Jones, Matthew R Helmus, and Katherine D McMahon. 2007. "Phylogenetic Ecology of the Freshwater Actinobacteria acI Lineage." *Applied and Environmental Microbiology* 73 (22) (November): 7169–76.

Oh, Seungdae, Alejandro Caro-Quintero, Despina Tsementzi, Natasha Deleon-Rodriguez, Chengwei Luo, Rachel Poretsky, and Konstantinos T Konstantinidis. 2011. "Metagenomic Insights into the Evolution, Function and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem." *Applied and Environmental Microbiology* 77 (17) (July 15): 6000–6011.

Pabst, Simone, Æ Nicole Scheifhacken, John Hesselschwerdt, and Æ Karl M Wantzen. 2008. "Leaf Litter Degradation in the Wave Impact Zone of a Pre-alpine Lake." *Hydrobiologia*: 117–131.

Pati, Amrita, Lenwood S Heath, Nikos C Kyrpides, and Natalia Ivanova. 2011. "ClaMS: A Classifier for Metagenomic Sequences." *Standards in Genomic Sciences* 5 (2) (November 30): 248–53.

Pedrós-Alió, C, and T D Brock. 1982. "Assessing Biomass and Production of Bacteria in Eutrophic Lake Mendota, Wisconsin." *Applied and Environmental Microbiology* 44 (1) (July): 203–18.

Pei, Jimin, and Nick V. Grishin. 2005. "COG3926 and COG5526 : A Tale of Two New Lysozyme-like Protein Families." *Protein Science* 14: 2574–2581.

Peura, Sari, Alexander Eiler, Stefan Bertilsson, Hannu Nyka, Marja Tiirola, and Roger I Jones. 2012. "Distinct and Diverse Anaerobic Bacterial Communities in Boreal Lakes Dominated by Candidate Division OD1": 1640–1652.

Pfennig, Norbert, and Germaine Cohen-Bazire. 1967. "Some Properties of the Green Bacterium Pelodictyon Clathratiforme." *Archiv Für Mikrobiologie* 59 (1-3): 226–236.

Poretsky, Rachel S, Shulei Sun, Xiaozhen Mou, and Mary Ann Moran. 2010. "Transporter Genes Expressed by Coastal Bacterioplankton in Response to Dissolved Organic Carbon." *Environmental Microbiology* 12 (3) (March): 616–27.

Pruesse, Elmar, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. 2007. "SILVA: a Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB." *Nucleic Acids Research* 35 (21) (January): 7188–96.

Raes, Jeroen, Jan O Korbel, Martin J Lercher, Christian von Mering, and Peer Bork. 2007. "Prediction of Effective Genome Size in Metagenomic Samples." *Genome Biology* 8 (1).

Sachse, A, D Babenzien, G Ginzel, Gelbrecth J, and C E W Steinberg. 2001. "Characterization of Dissolved Organic Carbon (DOC) in a Dystrophic Lake and an Adjacent Fen." *Biogeochemistry* 54 (3): 279–296.

Salcher, Michaela M, Jakob Pernthaler, and Thomas Posch. 2011. "Seasonal Bloom Dynamics and Ecophysiology of the Freshwater Sister Clade of SAR11 Bacteria 'That Rule the Waves' (LD12)." *The ISME Journal* 5 (8) (August): 1242–52.

Salcher, Michaela M, Thomas Posch, and Jakob Pernthaler. 2013. “In Situ Substrate Preferences of Abundant Bacterioplankton Populations in a Prealpine Freshwater Lake.” *ISME J* 7 (5) (May): 896–907.

Salcher, Michaela M., Jakob Pernthaler, and Thomas Posch. 2010. “Spatiotemporal Distribution and Activity Patterns of Bacteria from Three Phylogenetic Groups in an Oligomesotrophic Lake.” *Limnology and Oceanography* 55 (2): 846–856.

Savvichev, A S, I I Rusanov, D.Yu. Rogozin, E E Zakharova, O N Lunina, I A Bryantseva, S K Yusupov, N V Pimenov, A G Degermendzhi, and M V Ivanov. 2005. “Microbiological and Isotopic-Geochemical Investigations of Meromictic Lakes in Khakasia in Winter.” *Microbiology* 74 (4): 477–485.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan a Lesniewski, et al. 2009. “Introducing Mothur: Open-source, Platform-independent, Community-supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23) (December): 7537–41.

Soranno, P A, S L Hubler, S R Carpenter, and R C Lathrop. 2013. “Phosphorus Loads to Surface Waters : A Simple Model to Account for Spatial Pattern of Land Use.” *Ecology Society of America* 6 (3): 865–878.

Steffen, Morgan M, Zhou Li, T Chad Effler, Loren J Hauser, Gregory L Boyer, and Steven W Wilhelm. 2012. “Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents.” *PloS One* 7 (8) (January): e44002.

Swan, Brandon K, Ben Tupper, Alexander Sczyrba, Federico M Lauro, Manuel Martinez-Garcia, José M González, Haiwei Luo, et al. 2013. “Prevalent Genome Streamlining and Latitudinal Divergence of Planktonic Bacteria in the Surface Ocean.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (28) (July 9): 11463–8.

Tarr, M a, W Wang, T S Bianchi, and E Engelhaupt. 2001. “Mechanisms of Ammonia and Amino Acid Photoproduction from Aquatic Humic and Colloidal Matter.” *Water Research* 35 (15) (October): 3688–96.

Toolan, T, J D Wehr, and S Findlay. 1991. “Inorganic Phosphorus Stimulation of Bacterioplankton Production in a Meso-eutrophic Lake.” *Applied and Environmental Microbiology* 57 (7) (July): 2074–8.

Watanabe, Keiji, Nobuyuki Komatsu, Yuichi Ishii, and Masami Negishi. 2009. “Effective Isolation of Bacterioplankton Genus Polynucleobacter from Freshwater Environments Grown on Photochemically Degraded Dissolved Organic Matter.” *FEMS Microbiology Ecology* 67: 57–68.

Weiß, Michael, Zuzana Sýkorová, Sigisfredo Garnica, Kai Riess, Florent Martos, Cornelia Krause, Franz Oberwinkler, Robert Bauer, and Dirk Redecker. 2011. “Sebacinales Everywhere: Previously Overlooked Ubiquitous Fungal Endophytes.” *PloS One* 6 (2) (January): e16793.

Winnen, Brit, Rikki N Hvorup, and Milton H Saier. 2003. “The Tripartite Tricarboxylate Transporter (TTT) Family.” *Research in Microbiology* 154 (7) (September): 457–65.

Yannarell, A C, A D Kent, G H Lauster, T K Kratz, and E W Triplett. 2003. “Temporal Patterns in Bacterial Communities in Three Temperate Lakes of Different Trophic Status.” *Microbial Ecology* 46 (4) (November): 391–405.

Tables and Figures

Table 3-1: Sample characteristics

	Date	Reads	Million bp (mbp)	Avg Genome Size (mbp)	FunyBase hits per mbp	Taxon OID*	Contigs *	Million bp (mbp)*	Gene Count *
TBE03Jun09	June 3, 2009	1.90E+07	4515	3.41	116	3300000176	172147	120	2.60E+05
TBE18Aug09	August 18, 2009	2.40E+07	5310	3.81	119	3300000203	130632	82	1.80E+05
ME20Apr10	April 20, 2010	1.40E+07	3246	4.55	181	2199352003	54727	28	6.00E+04
ME15Jun10	June 15, 2010	2.10E+07	4891	2.92	179	2199352004	130296	71	1.60E+05
ME29Oct10	October 29, 2010	2.10E+07	4903	4.35	182	2199352005	127997	64	1.60E+05
TBcomposite	-	4.30E+07	9825	3.62	118	-	302779	202	4.40E+05
MEcomposite	-	5.60E+07	13040	3.39	180	-	313020	163	3.80E+05

*Assembled Data

Table 3-2: COGs overrepresented in either TB or ME and their Z-scores

COG	MEZ	TBZ	COG	MEZ	TBZ	COG	MEZ	TBZ
COG0463	3.41	9.66	**COG3864**	-0.28	2.02	**COG4397**	-0.60	1.07
COG3181	1.48	9.21	**COG0003**	-0.25	1.96	**COG0674**	-0.34	1.05
COG0234	1.26	7.91	**COG0741**	0.14	1.94	**COG0732**	-0.44	1.01
COG5323	-0.58	7.24	**COG0589**	-0.12	1.84	**COG5434**	-0.47	1.00
COG0553	0.83	6.98	**COG3328**	0.01	1.77	**COG1752**	-0.28	0.96
COG1216	1.82	6.63	**COG3628**	-0.14	1.73	**COG4186**	-0.17	0.96
COG3179	0.34	6.20	**COG5178**	-0.39	1.68	**COG2206**	-0.50	0.94
COG0629	1.43	6.17	**COG0602**	0.09	1.65	**COG0641**	-0.46	0.93
COG0863	1.29	6.15	**COG3637**	-0.49	1.63	**COG3587**	-0.45	0.92
COG5410	-0.34	5.94	**COG2865**	-0.33	1.55	**COG3245**	-0.33	0.89
COG2801	0.73	5.61	**COG5405**	-0.57	1.52	**COG1348**	-0.34	0.87
COG0535	0.47	5.27	**COG1961**	0.05	1.48	**COG1917**	-0.28	0.86
COG1484	-0.24	5.08	**COG4227**	-0.58	1.47	**COG4123**	-0.40	0.83
COG3926	-0.56	4.03	**COG4653**	-0.40	1.42	**COG1013**	-0.30	0.81
COG4584	-0.34	3.58	**COG4695**	-0.06	1.38	**COG3863**	-0.61	0.74
COG5377	-0.50	3.44	**COG0502**	-0.26	1.33	**COG4675**	-0.40	0.72
COG0790	-0.27	3.21	**COG3756**	-0.58	1.31	**COG3808**	3.31	0.69
COG5362	-0.32	3.16	**COG4643**	-0.61	1.29	**COG1775**	-0.61	0.67
COG0286	0.16	3.00	**COG4804**	-0.47	1.29	**COG3680**	-0.54	0.65
COG3772	-0.10	2.95	**COG0677**	-0.10	1.23	**COG2944**	-0.55	0.58
COG4420	-0.23	2.94	**COG4372**	-0.23	1.23	**COG0861**	2.58	0.41
COG2710	-0.14	2.83	**COG3093**	-0.14	1.22	**COG0538**	1.80	0.09
COG4570	-0.46	2.72	**COG3108**	-0.48	1.22	**COG4102**	1.68	0.07
COG0071	0.20	2.64	**COG1864**	-0.28	1.22	**COG1615**	2.23	0.05
COG3064	0.09	2.52	**COG3773**	-0.34	1.22	**COG1033**	1.12	-0.17

COG3378	-0.32	2.47	**COG1914**	-0.21	1.21	**COG2133**	0.97	-0.19
COG3497	0.00	2.39	**COG2963**	-0.18	1.20	**COG1520**	0.98	-0.23
COG4626	-0.05	2.35	**COG5281**	-0.42	1.19	**COG1748**	0.98	-0.24
COG0600	-0.05	2.33	**COG3128**	-0.08	1.19	**COG1122**	0.74	-0.29
COG1116	0.15	2.32	**COG4197**	-0.61	1.15	**COG0687**	1.10	-0.29
COG0715	0.01	2.18	**COG1541**	-0.19	1.12	**COG5464**	0.97	-0.44
COG4191	-0.07	2.17	**COG1793**	-0.11	1.12	**COG4636**	3.43	-0.45
COG4907	-0.33	2.14	**COG5519**	-0.61	1.10	**COG3509**	0.91	-0.50
COG1783	-0.05	2.07	**COG2202**	-0.21	1.08			

COG0003=Oxyanion-translocating ATPase; COG0071; COG0071=Molecular chaperone (small heat shock protein); COG0234=Co-chaperonin GroES (HSP10); COG0286=Type I restriction-modification system methyltransferase subunit; COG0463=Glycosyltranferases; COG0502=Biotin synthase and related enzymes; COG0535=Predicted Fe-S oxidoreductases; COG0538=Isocitrate dehydrogenases; COG0553=Superfamily II DNA/RNA helicases, SNF2 family; COG0589=Universal stress protein UspA; COG0600=ABC-type nitrate/sulfonate/bicarbonate transport system; COG0602=Organic Radical activating enzymes; COG0629=Single-stranded DNA-binding protein; COG0641=Arylsulfatase regulator; COG0674=Pyruvate:ferredoxin oxidoreductase; COG0677=UDP-N-acetyl-D-mannosaminuronate dehydrogenase; COG0687=Spermidine/putrescine-binding periplasmic protein; COG0715=ABC-type nitrate/sulfonate/bicarbonate transport systems; COG0732=Restriction endonuclease S subunits; COG0741=Soluble lytic murein transglycosylase; COG0790=FOG: TPR repeat, SEL1 subfamily; COG0861= Membrane protein TerC; COG0863=DNA modification methylase; COG1013=Pyruvate:ferredoxin oxidoreductase; COG1033=Predicted exporters of the RND superfamily; COG1216=Predicted glycosyltransferases; COG1116=ABC-type nitrate/sulfonate/bicarbonate transport system; COG1122= ABC-type cobalt transport system; COG1348=Nitrogenase subunit NifH; COG1484=DNA replication protein; COG1520=FOG: WD40-like repeat; COG1541=Coenzyme F390 synthetase; COG1615= Uncharacterized conserved protein; COG1748= Saccharopine dehydrogenase; COG1752=Predicted esterase; COG1775=Benzoyl-CoA reductase; COG1783= Phage terminase; COG1793= ATP-dependent DNA ligase; COG1864=DNA/RNA endonuclease G, NUC1; COG1914= Mn2+ and Fe2+ transporters; COG1917= Uncharacterized conserved protein; COG1961=Site-specific recombinases, DNA invertase Pin homologs; COG2133=Glucose/sorbosone dehydrogenases; COG2202=FOG: PAS/PAC domain; COG2206=HD-GYP domain; COG2710= Nitrogenase molybdenum-iron protein; COG2801=Transposase and inactivated derivatives; COG2865=Predicted transcriptional regulator; COG2944=Predicted transcriptional regulator; COG2963=Transposase and inactivated derivatives; COG3064=Membrane protein involved in colicin uptake; COG3093=Plasmid maintenance system antidote protein; COG3108=Uncharacterized protein conserved in bacteria; COG3128=Uncharacterized iron-regulated protein; COG3179=Uncharacterized iron-regulated protein; COG3181=Uncharacterized protein conserved in bacteria; COG3245=Cytochrome c5; COG3328=Transposase and inactivated derivatives; COG3378=Predicted ATPase; COG3497=Phage tail sheath protein FI; COG3509=Poly(3-hydroxybutyrate) depolymerase; COG3587=Restriction endonuclease; COG3638=Phage baseplate assembly protein W; COG3607=Opacity protein and related surface antigens; COG3680=Uncharacterized protein conserved in bacteria; COG3756=Uncharacterized protein conserved in bacteria; COG3772=Phage-related lysozyme; COG3773= Cell wall hydrolyses involved in spore germination; COG3863= distant relative of cell wall-associated hydrolases; COG3864=Uncharacterized protein conserved in bacteria; COG3926=Putative secretion activating protein; COG4102=Uncharacterized protein conserved in bacteria; COG4123=Predicted O-methyltransferase; COG4186=Predicted phosphoesterase or phosphohydrolase; COG4191=Signal transduction histidine kinase

regulating C4-dicarboxylate transport system; COG4197=Uncharacterized protein conserved in bacteria; COG4227= Antirestriction protein; COG4372=Uncharacterized protein conserved in bacteria; COG4397= Mu-like prophage major head subunit gpT; COG4420=Predicted membrane protein; COG4570=Holliday junction resolvase; COG4584=Transposase and inactivated derivatives; GOG4626=Phage terminase-like protein; COG4636=Uncharacterized protein conserved in cyanobacteria; COG4643=Uncharacterized protein conserved in bacteria; COG4653=Predicted phage phi-C31 gp36 major capsid-like protein; COG4675=Microcystin-dependent protein; COG4695=Phage-related protein; COG4804=Uncharacterized conserved protein; COG4907=Predicted membrane protein; COG5178=U5 snRNP spliceosome subunit; COG5281=Phage-related minor tail protein; COG5323=Uncharacterized conserved protein; COG5362=Phage-related terminase; COG5377=Phage-related protein, predicted endonuclease; COG5405=ATP-dependent protease HslVU (ClpYQ), peptidase subunit; COG5410=Uncharacterized protein conserved in bacteria; COG5434=Endopolygalacturonase; COG5464=Uncharacterized conserved protein; COG5519=Superfamily II helicase and inactivated derivatives.

Table 3-3: Abundance of assembled DOC transporters and PBEs per genome*

Function Category	ME	TB	% Difference
Carbohydrate Esterases	0.9	1.9	110
Glycoside Hydrolases	5.8	11.4	95
Polysaccharide Lyases	0.0	0.1	648
Sulfatases	1.2	1.2	0
Peptidases	16.4	19.4	18
Total PBEs	24.4	33.9	39
Carboxylic Acids	3.5	3.0	15
Carbohydrate	11.6	10.1	13
Nucleotide	1.2	0.6	53
Lipid	1.1	0.9	19
Amino Acids	16.0	21.2	33
Compatible solutes	1.0	1.5	40
Polyamines	3.8	1.5	60
Total DOC transporters	38.2	38.7	1

*Average genome sizes were calculated with reads using a modified version of Frank and Sorensen (2011).

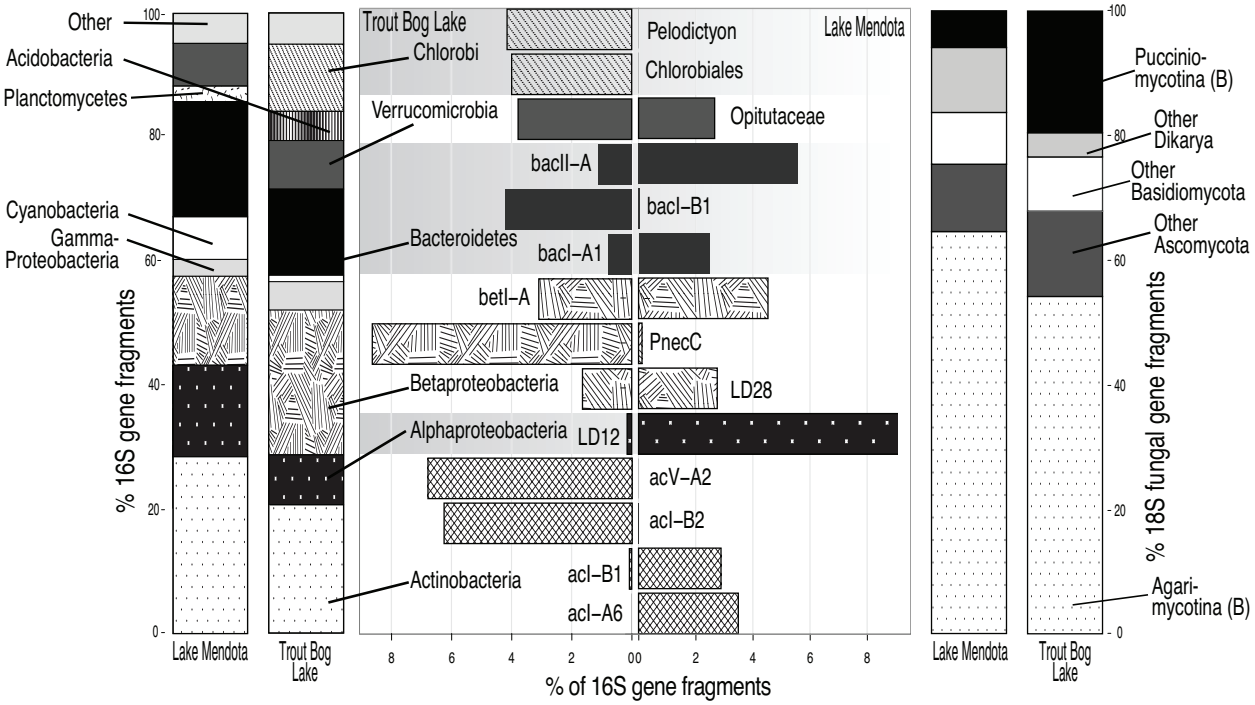

Figure 3-1: Community composition as estimated by SSU rRNA gene fragments that were recruited out of metagenomic datasets by BLASTn. From the phylum level (left), the communities were generally similar with the exception of different abundances of Chlorobi, Acidobacteria, and Cyanobacteria. At the tribe level (middle), common freshwater lineages make up the top 14 (combined % abundance) taxa, yet strong preference to a specific ecosystem is seen between related tribes. Analysis of fungal 18S (right) revealed that the ME and TB communities were generally similar at the subdivision level.

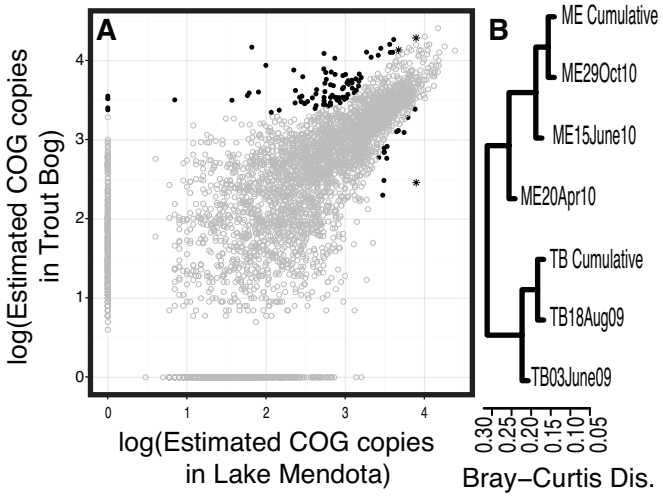

Figure 3-2: A) A scatter plot showing the relative abundance of each COG in ME and TB on a log scale. Black stars are COGs that are statistically overrepresented in either dataset ($p<0.1$); black circles are overrepresented COGs by defined criteria; gray points are COGs that were not considered for overrepresentation. B) A pairwise comparison of the functional content of each sample. The cluster was drawn with Bray-Curtis dissimilarity metrics, using the relative abundance of each COG annotation within a sample.

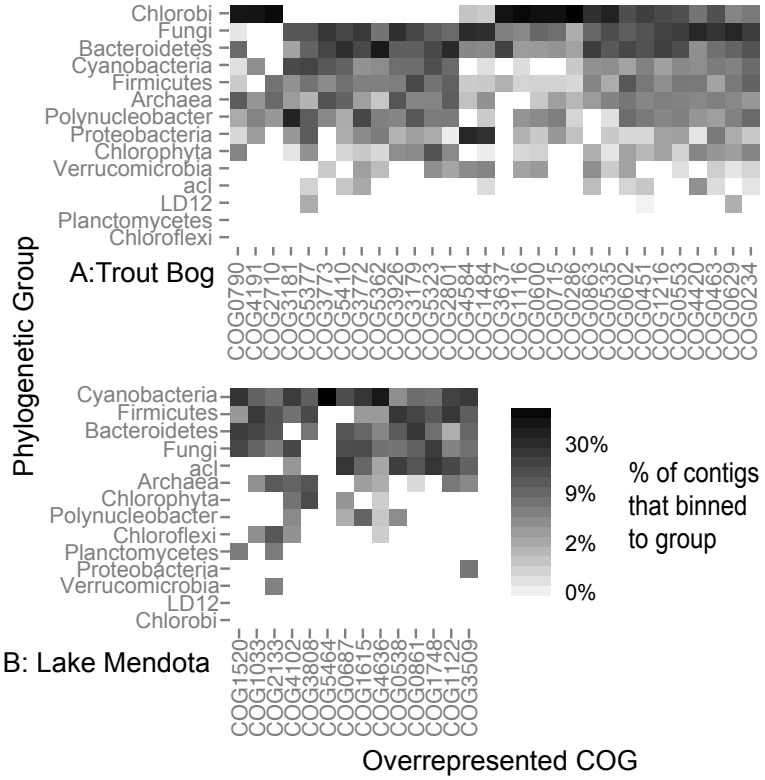

Figure 3-3: A) Phylogenic distribution of COGs overrepresented in TB. Binning was only performed on contigs that contained overrepresented COGs and were larger than 1kbp. B) Phylogenic distribution of COGs overrepresented in ME. Binning was only performed on contigs that contained overrepresented COGs and were larger than 1kbp.

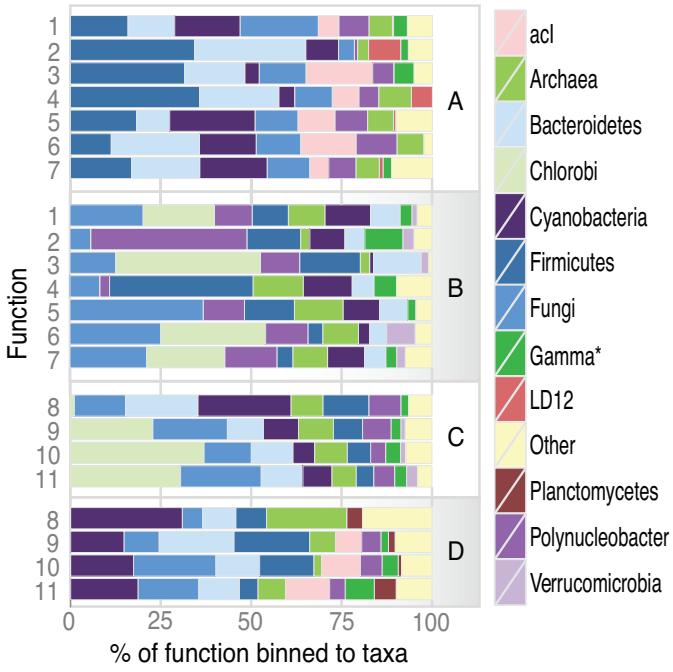

Figure 3-4: A) Best phylogenetic matches for contigs containing ABC-type DOC transporter COGs in ME. B) Best phylogenetic matches for contigs containing ABC-type DOC transporter COGs in TB. C) Best phylogenetic matches for contigs containing PBEs annotated with Enzyme Commission (EC) numbers in ME. D) Best phylogenetic matches for contigs containing PBEs annotated with EC numbers in TB. All percentages are weighted by contig read depths. Categories 1-7 are ABC-Type DOC transporters defined with COGs and 8-11 are PBEs defined with Enzyme Commission numbers. 1= carbohydrate; 2=carboxylic; 3=compatible solute; 4=lipid; 5=nucleotide; 6=polyamine; 7=amino acid; 8=sulfatase 9=peptidase; 10=glycoside hydrolases; 11=carbohydrate esterases.

Supplemental Methods

Pairwise Comparisons of COG Annotations

Xipe-Totec was used to identify the statistically overrepresented COGs for each sample to sample comparison. The online interface of XIPE was accessed (http://edwards.sdsu.edu/cgi-bin/xipe.cgi), and the following criteria were used: sample size = 5000, confidence level = 90%. Results for sample-to-sample comparisons are shown in Supplemental Table 5, and results from the composite Mendota and Trout Bog samples are shown in Supplemental Table 6.

Results from pairwise samples demonstrated that the one lake's samples typically had more overrepresented COGs when compared across lakes rather than within. Lake Mendota had fewer overrepresented COGs overall. The composite samples followed the same trends, and all but two COGs from this analysis were captured in the Z-Score overrepresentation criteria defined in the manuscript. These two COGs (COG1216 and COG0463), were thus considered overrepresented set for the manuscript analyses.

Supplemental Tables and Figures

Table S3-1: Site Characteristics

	Lake Mendota	Trout Bog
Area (km2)	39.9	0.011
Max Depth (m)	25.3	8
pH	8.5	4.8
DOC (mg/L)	6	25
TP (μg/L)	83	38
Trophic State	Eutrophic	Dystrophic

Table S3-2: Rank Abundance data for 16S reads from composite lake samples

Top 25 Taxa in each 16S read set

TB			ME		
Rank	%	Taxon	Rank	%	Taxon
1	11.5	Pnec	1	9.1	LD12
2	6.8	acV-A2	2	5.6	bacII-A
3	6.3	acI-B2	3	4.5	betI-A
4	4.2	bacI-b1	4	3.7	uncl. acI-A
5	4.1	Peolodictyon	5	3.5	acI-A6
6	4	uncl. Chlorobiales	6	2.9	acI-B1
7	3.8	Opitutaceae	7	2.8	betIV-A
8	3.1	betI-A	8	2.7	Opitutaceae
9	2.1	Verrucomicrobiales sp	9	2.6	Dolichospermum sp
10	2.1	Holophagaceae sp	10	2.5	bacI-A1
11	1.7	uncl. bacI-A	11	2.5	acI-A7
12	1.7	Geothrix sp	12	1.9	uncl. acI
13	1.6	betIV-A	13	1.8	Illuma-A2
14	1.6	acI-B3	14	1.6	bacIII-A
15	1.5	uncl. acI-B	15	1.4	uncl. bacV
16	1.3	Chlorobaculum sp	16	1.2	Pnec
17	1.3	acV-A1	17	1.2	Luteolibacter sp
18	1.3	Acetobacteracaea sp	18	1.1	Xip-B1
19	1.2	bacVI-A	19	1.1	SOGA31 sp (Chloroflexi)
20	1.2	Crenothrix sp	20	1.1	alfIV-B
21	1.2	acI-B4	21	1.1	uncl. Nostocaceae
22	1.1	Sinobacteraceae sp	22	1.1	uncl. bacVI-B
23	1.1	bacII-A	23	1	uncl. acTH1-A1
24	1.1	Chlorobiaceae sp	24	1	betI-B
25	1	uncl. Methylotenera	25	0.9	uncl. bacI

Table S3-3: COGs present in ME, absent in TB

COG5235	COG2082	COG5413	COG3966	COG2209	COG1498	COG1632	COG5187	COG5128	COG1634
COG4327	COG2164	COG5437	COG5041	COG3856	COG4940	COG1031	COG1877	COG5202	COG2443
COG1133	COG5575	COG5522	COG5258	COG2238	COG5169	COG1499	COG1911	COG1712	COG3918
COG1140	COG2310	COG5044	COG5403	COG3743	COG5252	COG1631	COG4991	COG3351	COG4088
COG4764	COG4370	COG2389	COG0710	COG1578	COG1471	COG3210	COG1394	COG4594	COG4110
COG4240	COG1805	COG3457	COG4385	COG4830	COG2139	COG4667	COG2214	COG4903	COG4282
COG4336	COG5564	COG3861	COG4335	COG2157	COG5647	COG4900	COG3725	COG5087	COG4904
COG4729	COG1903	COG4360	COG5493	COG3037	COG0478	COG5035	COG4892	COG5157	COG5058
COG5274	COG4902	COG2988	COG1155	COG5554	COG2915	COG5052	COG2261	COG5513	COG5103
COG0383	COG4308	COG1713	COG2312	COG1971	COG4247	COG5171	COG3516	COG1367	COG5231
COG2180	COG4872	COG4112	COG1515	COG4691	COG5491	COG5228	COG3901	COG2219	COG5232
COG5077	COG5011	COG4290	COG2979	COG5163	COG3422	COG2167	COG4758	COG2770	COG5260
COG0371	COG4402	COG1563	COG5253	COG5221	COG3848	COG2733	COG4851	COG3559	COG5439
COG5059	COG2035	COG1808	COG5420	COG2163	COG4606	COG4322	COG5076	COG5095	COG5616
COG5454	COG4241	COG3977	COG1103	COG4125	COG4813	COG5293	COG5084	COG5143	COG1292
COG4756	COG4492	COG2241	COG4352	COG5063	COG1431	COG5444	COG5217	COG5653	COG1339
COG4250	COG3689	COG5273	COG1084	COG3018	COG3423	COG1717	COG1779	COG2024	COG1379
COG4552	COG2099	COG1726	COG1939	COG3872	COG3492	COG3045	COG2058	COG2042	COG1628
COG5277	COG1900	COG4877	COG4671	COG4803	COG1537	COG3603	COG2946	COG2078	COG1727
COG3781	COG5141	COG3379	COG4759	COG5051	COG1759	COG3853	COG3767	COG3738	COG2092
COG4576	COG1241	COG3919	COG5069	COG1890	COG1769	COG5325	COG3804	COG4493	COG3498
COG2866	COG5534	COG5549	COG2123	COG5108	COG2126	COG5387	COG4891	COG5144	COG5068
COG1572	COG1336	COG1010	COG3907	COG4098	COG2383	COG1601	COG5391	COG5165	COG1374
COG0585	COG2144	COG1156	COG5024	COG4718	COG4880	COG2147	COG5541	COG1730	COG1456
COG1115	COG1915	COG5438	COG5030	COG4063	COG5092	COG3843	COG1288	COG1958	COG1998
COG3001	COG2052	COG4337	COG1852	COG5081	COG1568	COG4061	COG1823	COG3044	COG2952
COG4477	COG2720	COG2880	COG4274	COG1347	COG2051	COG4741	COG3650	COG4525	COG3855
COG4100	COG5626	COG1604	COG1361	COG1552	COG2125	COG5118	COG3720	COG5061	COG4901
COG1791	COG3310	COG1930	COG3881	COG4171	COG2231	COG5150	COG5198	COG5477	COG5043
COG4935	COG4454	COG5034	COG4574	COG4269	COG4685	COG5259	COG2451	COG5591	COG5159
COG3010	COG4829	COG0819	COG4825	COG4298	COG2150	COG5660	COG3667	COG1097	COG5200
COG4868	COG2839	COG1503	COG5210	COG5032	COG3368	COG1527	COG5050	COG1867	COG5583
COG4229	COG4719	COG5498	COG5383	COG0786	COG3805	COG1889	COG5100	COG4086	COG5638
COG2073	COG5593	COG3554	COG1755	COG3326	COG3896	COG3048	COG1693	COG4328	COG1258
COG1224	COG4929	COG5119	COG1938	COG4238	COG4412	COG3714	COG2746	COG5017	COG1358
COG1659	COG1963	COG4682	COG2213	COG4297	COG2053	COG4592	COG3433	COG5156	COG2106
COG5256	COG3793	COG3105	COG2810	COG1093	COG3154	COG5045	COG4652	COG5173	COG2978
COG5426	COG0699	COG3232	COG5170	COG1163	COG3297	COG2097	COG4818	COG5209	COG3058
COG5031	COG3860	COG3271	COG1976	COG2004	COG4388	COG4387	COG5056	COG5536	COG3149
COG1780	COG5474	COG4031	COG2111	COG2136	COG4688	COG5071	COG5120	COG5563	COG3449
							COG3670	COG3214	COG4073

Table S3-4: COGs present in TB, absent in ME

COG4643	COG2162	COG1679	COG0549	COG5346	COG4357	COG3916	COG3218	COG4936	COG4905
COG4197	COG1332	COG5295	COG1263	COG4237	COG5338	COG4068	COG3838	COG4997	COG5472
COG3863	COG1583	COG3285	COG5570	COG5624	COG2739	COG4578	COG4223	COG1017	COG1683
COG1775	COG1553	COG4311	COG0827	COG1746	COG3562	COG2247	COG5079	COG2043	COG1798
COG3567	COG4655	COG4917	COG2410	COG4379	COG3592	COG2342	COG1415	COG2155	COG2019
COG2359	COG1342	COG1656	COG5298	COG4754	COG4058	COG5181	COG3068	COG3301	COG2056
COG4285	COG5479	COG4801	COG3837	COG5425	COG2874	COG5452	COG3072	COG4085	COG3814
COG1945	COG1908	COG5516	COG5109	COG5555	COG3610	COG3652	COG3143	COG4425	COG3945
COG3376	COG2074	COG1039	COG2983	COG3260	COG3892	COG4261	COG3535	COG4545	COG4849
COG4734	COG3025	COG5615	COG4941	COG5042	COG3405	COG4765	COG4894	COG4861	COG4871
COG4280	COG5302	COG3302	COG1718	COG1776	COG3683	COG5566	COG5442	COG5199	COG4922
COG5565	COG5457	COG4318	COG2134	COG4520	COG4884	COG1771	COG5465	COG2411	COG5014
COG1144	COG3534	COG4926	COG2933	COG1114	COG1812	COG3471	COG3701	COG3413	COG5186
COG4986	COG5188	COG3125	COG4193	COG2240	COG4380	COG3698	COG3816	COG4292	COG5343
COG3305	COG3692	COG3948	COG2861	COG4820	COG1549	COG2511	COG4160	COG4845	COG5430
COG4887	COG4722	COG5237	COG3137	COG4540	COG5429	COG3231	COG4809	COG5154	COG2704
COG2990	COG4384	COG2837	COG1710	COG5269	COG5622	COG3739	COG5494	COG5211	COG3151
COG0011	COG3312	COG3065	COG3488	COG4393	COG3426	COG3827	COG2851	COG5447	COG3583
COG2145	COG1731	COG4669	COG3819	COG3611	COG5308	COG4192	COG3570	COG1221	COG3942
COG1969	COG4573	COG3580	COG3824	COG4726	COG3935	COG5366	COG3646	COG1542	COG4752
COG1993	COG2375	COG3866	COG4119	COG5291	COG4211	COG5633	COG4334	COG3645	COG5381
COG1102	COG5489	COG4032	COG4906	COG1152	COG4325	COG1273	COG4767	COG3697	COG1098
COG2094	COG4496	COG4690	COG5400	COG3121	COG4837	COG1768	COG4859	COG3826	COG3251
COG1382	COG4386	COG3180	COG3594	COG4517	COG5342	COG1829	COG5443	COG4056	COG3946
COG3524	COG2517	COG5283	COG3960	COG4808	COG5414	COG3519	COG1150	COG4101	COG4271
COG3644	COG3619	COG5384	COG5270	COG5055	COG2604	COG4739	COG1464	COG5242	COG4342
COG1337	COG4458	COG4712	COG3904	COG2116	COG3056	COG5321	COG3412	COG5328	COG4461
COG3398	COG3070	COG3261	COG4914	COG5083	COG5462	COG5568	COG3543	COG1184	COG4713
COG4924	COG4834	COG3898	COG3952	COG3778	COG1786	COG2879	COG3709	COG1323	COG4776
COG5634	COG2357	COG4508	COG3240	COG5180	COG1787	COG3821	COG3890	COG1378	COG5025
COG3155	COG1910	COG3817	COG3799	COG2190	COG3626	COG4449	COG4008	COG3835	COG5218
COG2830	COG3780	COG3109	COG4537	COG3512	COG4316	COG5026	COG4093	COG3840	COG3589
COG1742	COG1639	COG3888	COG5107	COG1145	COG5385	COG5509	COG4296	COG3915	COG3106
COG5448	COG5514	COG3921	COG5585	COG2362	COG5631	COG5571	COG4925	COG4572	
COG4277	COG3704	COG0650	COG3665	COG1124	COG3304	COG1356	COG4978	COG1831	
COG1421	COG5296	COG3581	COG4266	COG1275	COG3760	COG2413	COG5317	COG1863	
COG1567	COG4396	COG3489	COG5628	COG3953	COG4027	COG4130	COG1455	COG3300	
COG3416	COG2872	COG1153	COG5562	COG3342	COG5532	COG5330	COG3196	COG4321	
COG2406	COG3131	COG3331	COG5484	COG1817	COG1318	COG5595	COG4286	COG4331	
COG2168	COG5397	COG3504	COG1406	COG5266	COG1550	COG3787	COG4314	COG4607	
COG2923	COG3019	COG5621	COG3765	COG2966	COG3076	COG4701	COG4792	COG4895	

Table S3-5: Pairwise sample overrepresentation using Xipe-Totec

COGs overrepresented in samples in rows, when compared pairwise to columns.

	ME20Apr2010	ME15Jun2010	ME29Oct2010	TB03Jun2009	TB18Aug2009
ME20Apr2010	-	COG0328, COG0366, COG0395, COG0507, COG1012, COG1175, COG1960, COG2036, COG2826, COG3181, COG5013, COG5271, COG5560	COG0328, COG0366, COG0507, COG1061, COG1100, COG2036, COG5271, COG5560	COG0328, COG0366, COG0507, COG1012, COG5560	COG0085, COG0086, COG0210, COG0215, COG0328, COG0395, COG0459, COG0507, COG0515, COG0524, COG1012, COG1100, COG1131, COG1175, COG1233, COG1609, COG1615, COG1653, COG1960, COG2124, COG2409, COG5560
ME15Jun2010	COG0451, COG0859, COG1262, COG1520, COG3119, COG4636	-	COG0515	COG0515, COG3119, COG4636	COG0515, COG1520, COG3119, COG4636
ME29Oct2010	COG0675, COG1020, COG1538, COG3464, COG4636,COG5464	COG0507, COG0675, COG3464	-	COG0675, COG3464, COG4636, COG5464	COG0675, COG3464, COG4636
TB03Jun2009	COG0234, COG0417, COG0463, COG0535, COG0553, COG0629, COG0749, COG0859, COG0863, COG1216, COG1538, COG3179, COG3181, COG3926, COG4570, COG4626, COG5323, COG5377, COG5410)	COG0234, COG0553, COG0629, COG3179, COG3181, COG3926, COG4420, COG4570, COG5323, COG5377, COG5410	COG0234, COG0417, COG0438, COG0463, COG0535, COG0553, COG0629, COG0749, COG0863, COG1061, COG1216, COG1484, COG2826, COG3064, COG3179, COG3181, COG3378, COG3926, COG4420, CGO4570, COG4626, COG4907, COG5178, COG5323, COG5362, COG5377, COG5410, COG5434	-	N/A
TB18Aug2009	COG0003, COG0071, COG0234, COG0286, COG0358, COG0399, COG0438, COG0451, COG0463, COG0502, COG0535, COG0589, COG0600, COG0642, COG0715, COG0749, COG0776, COG0790, COG0845, COG0859, COG0863, COG1032, COG1116, COG1136, COG1216, COG1357, COG1484, COG1487, COG1538, COG2204, COG2206, COG2227, COG2710, COG2801, COG2865, COG2885, COG3179, COG3497, COG3637, COG3772, COG3926, COG4191, COG4584, CO4591, COG5323, COG5410	COG0003, COG0234, COG0286, COG0535, COG0553, COG0589, COG0600, COG0790, COG1032, COG1116, COG1484, COG1538, COG2199, COG2206, COG2710, COG2801, COG2865, COG3179, COG3181, COG3378, COG3587, COG3637, COG3926, COG4191, COG4584, COG4804, COG5323, COG5362, COG5377, COG5410	COG0003, COG0234, COG0286, COG0399, COG0417, COG0438, COG0451, COG0463, COG0502, COG0535, COG0589, COG0600, COG0642, COG0715, COG0749, COG0790,COG0845, COG0863, COG1032, COG1061, COG1216, COG1484, COG1538, COG1775, COG2199, COG2227, COG2710, COG2801, COG2865, COG3179, COG3181, COG3378, COG3587, COG3637, COG3772, COG3926, COG4191, COG4227, COG4570, COG4584, COG4884, COG4974, COG5323, COG5362, COG5377, COG5410	COG0003, COG0286, COG0600, COG0642, COG0790, COG1032, COG1116, COG1357, COG1484, COG1775, COG2199, COG2206, COG2710, COG2865, COG2886, COG3245, COG3587, COG3637, COG4191, COG4584, COG4804	-

Table S3-6: Pairwise overrepresentation for composite metagenomic samples using Xipe-Totec

COGs overrepresented in samples in rows, when compared pairwise to	ME	TB
ME	-	COG4636
TB	COG0234, COG0463, COG0535, COG0553, COG0629, COG0790, COG0863, COG1216, COG1484, COG2710, COG2801, COG3179, COG3181, COG3926, COG4420, COG4570, COG4584, COG5323, COG5362, COG5377, COG5410	-

Future Work

The prologues in the data chapters state that Chapter 2 and Chapter 3 are both in preparation for submittal. In the coming weeks and months, these chapters will under go review and edits until they are accepted for publication.

I will be continuing to be advised and mentored by Dr. McMahon during the pursuit of my Ph.D. in environmental engineering at the University of Wisconsin – Madison. My research will continue to be focused on the role of bacterial communities and populations in the biogeochemical cycling of organic matter, using a blend of molecular and computational techniques. Upon submission on the two data chapters in this thesis, I intend to focus my efforts on the four fronts described below.

1. **Determination of a core community in humic bog lakes.** Using the same 16S tag dataset that was analyzed on the community level in Chapter 2, I intend to define a core set of populations that behave in predictable manners across the set of 7 humic lakes. Progress on this project has already been made, but more effort will be placed into finding co-occurrence patterns between taxa and how these core populations behave dynamically upon the influence of environmental characteristics.
2. **Identification of catchment area influence on specific aquatic taxa.** A collaboration began this summer with a Ph.D. student (Samantha Oliver) in the Stanley lab at the Center for Limnology, and samples from approximately 30 small lakes have been collected. Catchment and lake characteristics have been documented for each lake, and I will attempt to correlate these characteristics with specific taxa. Samantha Oliver designed the sampling strategy and collected the samples. My role in the collaboration was to collect bacterial biomass from the samples, and will subsequently be to use molecular techniques to characterize the samples from each lake for bacterial community structure.
3. **Taxon specific degradation of algal detritus, leaf litter, and cellulose**. Using data from single-cell amplified genomes, I have hypothesized that individuals from Verrucomicrobia and Bacteroidetes phyla have similar but distinct substrate preferences but are both active in the decomposition of polysaccharides and proteins. This summer, I designed addition experiments that will be used to target the role of these two phyla in the degradation of pure particulate substrates (cellulose), particulate leaf litter, and algal detritus. The cellulose addition has been recently been completed using bacterial communities from Trout Bog, and preparations for the leaf litter and algal detritus additions are being made.
4. **Fine-scale temporal trends in functional content in Trout Bog and Lake Mendota.** Approximately 90 metagenomic libraries from Trout Bog and 90 metagenomic libraries from Lake Mendota have been constructed. Using the insights gained from Chapter 3, we

will investigate the temporal dynamics of evidence for substrate availability. In addition, the role of hypolimnetic communities will be included in this analysis.