

**Genomics as a Lens into the Evolution and Ecology of Freshwater  
Microbes**

**Sarah Stevens  
Preliminary Proposal**

**Preliminary Presentation  
June 10<sup>th</sup>, 2014  
1 pm  
MSB 5503**

## **Abstract**

Research of freshwater microbes has been transformed by the availability of metagenomes, from which genomes are obtained using differential coverage, and by single amplified genomes (SAGs). Before these advances, genomic research was limited to easily cultivatable microbes. Now, genomes binned from metagenomes (GFMs) and SAGs allow for unprecedented understanding of the genomics of lake communities. We are applying these innovative techniques to 192 metagenomes from two very different freshwater lakes. Our GFMs and SAGs, in conjunction with our long-term metagenomic time series, can be used to track how populations evolve and change over time. In addition, these genomes allow for the study of metabolic potential, providing insight into the functions these microbes may perform in nature. We can also study genome streamlining in lakes, which is a lifestyle common in marine microbes.

## **Specific Aims**

Overarching Goal: To use genomics to learn about the ecology and evolution of microbes in lakes, specifically how populations change over time and the metabolic potential and genomic features of lake bacteria.

### **AIM 1:**

Bacterial evolution in nature has not previously been well studied because most microbes cannot be cultivated and it is nearly impossible to track natural populations on a whole genome scale. Most such efforts have compared natural populations using whole genomes from cultivated organisms isolated from a single kind of environment. There have also been a few studies that have mapped metagenomes to genomes from cultivated organisms. However, neither of these methods considers the uncultivated majority or attempts to look at the bacterial populations in a single site on a broader scale. In lakes, many of the abundant microbes have yet to be cultivated and isolated, such as the ubiquitous and abundant *act* lineage of actinobacteria. Using our metagenomic time series, which spans multiple years in two very different lakes (Trout Bog and Lake Mendota), we can use coverage-based binning and sequence features to retrieve reference genomes that are directly relevant to our dataset and not necessarily from cultivatable lineages. We also have single amplified genomes (SAGs), representing a variety of major

freshwater lineages, many of which were recovered from Lake Mendota. We can map metagenomic reads back onto these genomes from metagenomes (GFMs) and onto the SAGs and in order to study the population structure and dynamics over time.

I plan to focus specifically on two cosmopolitan freshwater groups: the LD12 tribe, which is the only tribe of the *alfV* lineage that is found in lakes, and the *acl* lineage. In our controlled freshwater taxonomy vocabulary, a lineage is defined as a monophyletic group by phylogenetics within a phylum; whereas a tribe is monophyletic by phylogenetics and has  $\geq 97\%$  16S rRNA gene sequence identity (Newton *et al.*, 2011). My approach is novel because coverage based binning techniques have only recently been developed. Our metagenomic time series is also unprecedented in its length and coverage, including a total of 192 samples sequenced using Illumina HiSeq technology. To date, I have worked with collaborators at the Joint Genome Institute (JGI) on the evolutionary dynamics, specifically single nucleotide polymorphisms (SNPs) and gene gain and loss, in a smaller set of manually binned genomes, for which there is a manuscript currently in review.

**AIM 1: To determine the evolutionary dynamics of sequence discrete populations of bacteria in freshwater lakes, we will map reads from our metagenomic time series to composite GFMs and SAGs and examine abundance patterns, SNP patterns, gene gain and loss, and recombination.**

*H1.1 GFMs from lakes with more stable biogeochemical conditions over time have less variation in relative abundance.*

*H1.2: As predicted due to their low diversity and low rates of recombination, the LD12 tribe experiences genome-wide sweeps over time.*

*H1.3: The *acl* lineage, which has much greater diversity in comparison with the LD12 tribe, undergoes gene-specific sweeps over time and has higher rates of recombination.*

*H1.4: Each population undergoes gene-specific or genome-wide sweeps when under selective pressure, depending on their rates of recombination.*

## **AIM 2:**

There is little known about the functions these freshwater bacteria are performing in the environment. For example, without cultivation, experiments cannot be performed to directly evaluate substrate uptake. Microautoradiography and fluorescence *in situ* hybridization (MAR-FISH) can be used to observe which cells are taking up a radio-labeled substrate but it is a laborious and expensive

experimental method. In order to reduce the number of substrates to test and to focus on key candidate lineages there must be some information known about metabolic potentials. Gene and pathway annotations/predictions in the GFMs and SAGs will provide clues for such functional potentials in specific key lineages. We can use the evolutionary patterns and population dynamics found above to target functions applying evolutionary pressures to microbes in the environment. Our lab has previously been involved in looking at functional capabilities of the *acl*-actinobacteria lineage using SAGs (Garcia *et al.*, 2013; Ghylis *et al.*, *in review*). We can also search for temporal correlations between nutrient conditions in the lakes and relative abundances of the sequence discrete populations determined in Aim 1. Learning more about the metabolic functions that specific groups of microbes are performing in the environment can help us to better understand the biogeochemical cycling being carried out.

**AIM 2: To generate hypotheses about which bacterial groups are performing specific metabolic functions in the community, we will characterize the functional potential of bacteria using GFMs and SAGs.**

### **AIM 3:**

Genome streamlining has been observed in freshwater and marine settings but little is understood about the features that make this lifestyle successful in aquatic environments. As previously mentioned, freshwater systems have very few reference genomes and many of those previously studied were cultivated. It has also been suggested that previous difficulty in culturing these bacteria may be due to their streamlined nature (Giovannoni *et al.*, 2005). Furthermore, the genomes from cultivated bacteria may not be representative of those microbes that are most common or abundant. To learn if streamlined genomes are common among freshwater microbes, we need to analyze the genomes from uncultivated organisms. We propose to characterize the genome features of the same GFMs and SAGs from Aims 1 and 2 to find evidence and possible mechanisms of genome streamlining. This substantial data set provides the opportunity to compare the whole genomes, many which may be streamlined. We can also observe if streamlined genomes are more common among abundant bacteria in the lake, and discover features that are different between streamlined and non-streamlined genomes from the same lake. Analysis of genome streamlining will help us to learn how important genome reduction is for successful populations in aquatic environments. This work will improve our understanding of genomic streamlining

as a microbial lifestyle and how it functions in a community setting.

**AIM 3: To determine if genome streamlining is a general characteristic of abundant freshwater bacteria, and if it is predicted by certain genomic features that I will search for in our GFMs and SAGs.**

*H3.1: As genome streamlining increases, the portion of genes allowing for a diversity of carbon substrate utilization decreases*

*H3.2: Genome streamlining correlates negatively with growth rate.*

*H3.3: As genome streamlining increases, the portion of genes associated with motility or signal transduction decreases.*

*H3.4: Streamlined genomes are traits of highly successful, abundant freshwater microbes.*

## **Background and Significance**

The identification and study of microbes has long depended on our ability to culture them. However many microbes are not amenable to growing in the lab environment (Amann *et al.*, 1995). The development of molecular techniques, such as high-throughput sequencing and genomics, has enabled us to start studying which microbes reside in an environment and what they are doing. Many approaches have allowed scientists to peek into the uncultivated world, such as 16S rRNA gene sequencing, metagenomics, and sequencing single cells. Each method has different advantages and disadvantages, but all can be used to answer specific questions about the microbes in an environment.

In freshwater lakes, the 16S rRNA gene has often been used to determine which microbes are in the lake. This method has been used to define the taxonomy of bacteria commonly found in lakes and to develop a highly curated 16S database (Newton *et al.*, 2011). While this approach, with a good reference database, can help to answer the question of which microbes inhabit the lake, it also does not inform what genomic features or genes these microbes may have. As this method uses a conserved marker, it does not capture more recent evolutionary trends such as horizontal gene transfer and mutations in genes under evolutionary pressures.

Another tool that can yield genomic information about the uncultured microbes is metagenomics (Gilbert & Dupont, 2011). This method involves extracting all DNA from lake water samples and sequencing it. One limitation of this technique is that it is difficult to track back the reads or even the assembled contigs to individual organisms or closely related populations. A recently developed technique

uses differential coverage to separate out contigs that belong to the same organism (Wrighton *et al.*, 2012; Albertsen *et al.*, 2013). This method requires two or more different metagenomes from the same ecosystem. One such method was used to separate out 49 draft genomes from an acetate amended aquifer using three metagenomes collected from different times (Wrighton *et al.*, 2012). From these genomes, metabolic potential was inferred from annotated genes. Where tetranucleotide frequency was used first in Wrighton *et al.*, another such study used the differential coverage as the first metric for binning genomes from an activated sludge bioreactor (Albertsen *et al.*, 2013). In this study the bins were further refined using tetranucleotide frequency, GC content, length, and single copy gene analysis. While both of the previous examples represent less complex microbial communities than those found in lakes, the lake community is a good next test set as compared to complex communities such as those found in soil. However, one must keep in mind that with more complex systems, assembly becomes more difficulty and rare populations may be missed due to insufficient sequencing coverage.

Another recently developed technique for investigating the genomes of uncultured organisms is whole genome amplification from single cells. Each cell is sorted by fluorescent activated cell sorting, lysed, amplified with whole genome amplification, sequenced and then assembled. From this process, single amplified genomes (SAGs) are created. The genes are then predicted and annotated using standard computational genome analysis methods (Woyke *et al.*, 2009). This method was applied to produce 201 SAGs from never before sequenced lineages (Rinke *et al.*, 2013). These phyla were known to exist in nature due to 16S sequencing but had never been isolated. These genomes markedly expanded the reference database to include these phyla and resolved broader evolutionary patterns of related phyla.

SAGs can also be used in conjunction with metagenomic reads to track populations. Two SAGs from marine flavobacteria were used as references for metagenome sequences from the Global Ocean Sampling (GOS) expedition (Woyke *et al.*, 2009). Each read from the metagenome is checked to see if it matches against the reference and is considered recruited if it matches above a specified identity threshold. These SAGs recruited reads from the GOS expedition much better than cultured marine flavobacteria. The authors recruited reads from different GOS sites and saw that the populations of flavobacteria closely related to these SAGs were likely dispersed by a known ocean current. This study is

an example of how important it is to consider uncultured organisms and to have genome references that are relevant to the environment in question. It is also an example of tracking populations using SAGs as references for metagenomes.

One important limitation with SAGs is that the whole genome amplification process has random bias and can miss whole sections of the genomes. A common method to estimate how complete these genomes are uses single copy conserved genes shared among bacteria or a particular bacterial lineage (Rinke *et al.*, 2013; Albertsen *et al.*, 2013; Garcia *et al.*, 2013).

Bacterial diversification and speciation is of great interest in microbial ecology (Cordero & Polz, 2014). The ecotype model is one popular theory proposed to explain bacterial speciation in nature (Cohan & Perry, 2007). However, support for this theory came only from models and inference since direct observations in support of it were not available (Chan *et al.*, *in review*). Our collaborators at JGI have manually binned several genomes, based on tetranucleotide frequency, phylogenetic gene distribution, and differential coverage, from our Trout Bog hypolimnion metagenomes (Chan *et al.*, *in review*). They then mapped the metagenome reads back to the genomes and identified single nucleotide polymorphisms (SNPs). All of the SNPs in the genomes they explored went to fixation, genome-wide over three years. This provides direct evidence for the ecotype model of speciation, which predicts genome-wide sweeps due to low rates of recombination.

Little is known about what functions bacteria are performing in nature. This is especially true for microbes that cannot be cultured, since experiments cannot be performed to test for functions. To find possible bacterial functions in nature, there have been a number of studies characterizing metabolic potential from annotated genomes. One influential paper characterizes the functional potential of 32 marine isolates from the lineage *Roseobacter* and was authored by a former member of the McMahon Lab (Newton *et al.*, 2010). This study systematically characterizes these genomes and their functional potential. A similar functional characterization has been done for our SAGs from the *acl* lineage (Garcia *et al.*, 2013; Ghylis *et al.*, *in review*).

Experimental evidence for lineage-specific traits has also been generated for some freshwater groups. To directly test for substrate uptake some have used microautoradiography and fluorescence *in situ* hybridization (MAR-FISH) to evaluate incorporation of 14 radiolabeled dissolved organic compounds

in 30 common freshwater bacterial groups at different levels of taxonomic resolution (Salcher *et al.*, 2013). This represents great strides in what is known about bacterial functions in the lake but is limited by the small number of substrates could be tested and by known FISH-defined phylogenetic groups.

Genome streamlining may be a key factor in why many of the microbes in the lake are so difficult to culture (Giovannoni *et al.*, 2014) and may also contribute to the success of common and abundant lineages. This feature has also been observed in endosymbionts and in the ubiquitous marine bacterium SAR11 (Grote *et al.*, 2012; Viklund *et al.*, 2012; Giovannoni *et al.*, 2005). One of the first freshwater bacterial genomes to be published was that of the streamlined *aci-B1* tribe (Garcia *et al.*, 2013). In another study, which published a number of genomes from freshwater isolates, the genomes were compared and a variety of lifestyles were predicted based on growth rate, carbon usage, and signal transduction and motility genes (Livermore *et al.*, 2013).

## **Preliminary Results**

### *Genomes from Metagenomes (GFMs)*

I worked with collaborators while at JGI who were developing a program called Metabat to bin genomes from metagenomes. Their program uses differential coverage across the 46 metagenomes from the epilimnion of Trout Bog, the 47 metagenomes from the hypolimnion of Trout Bog, and the 97 metagenomes from the epilimnion of Lake Mendota to bin together contigs. It also takes into account sequence features such as k-mer frequency when binning. The bins are statistically selected by the program and require no manual intervention to choose bins as is done with other methods. From this program, I recovered 87 GFMs from Trout Bog epilimnion, 167 GFMs from Trout Bog hypolimnion, and 502 GFMs from Lake Mendota. I then filtered the output to include only genomes which were 50% complete by single copy conserved genes found in 90% of all bacteria, and for genomes where 90% of those single copy genes were unique. After this filtering, we have 36, 70, and 104 GFMs from Trout Bog epilimnion, Trout Bog hypolimnion, and Lake Mendota epilimnion respectively.

In order to classify these genomes, I used a program called Phylosift (Darling *et al.*, 2014). It finds 37 conserved evolutionary marker genes using hidden markov models, aligns them, and gives a probability of classification for each marker gene. It is designed to be used on metagenomes and give the



user an idea of which organisms are in a metagenome by each marker. I created a parsing program in python to take the probabilities and classifications for each marker gene and interpret their results as from one organism. My program gives a classification based on a probability cutoff and a percent matching cutoff. At each level of Linnaean classification, it removes any hits below the probability cutoff and then if the hits left match above the percent matching threshold, it will save that classification and proceed to the next Linnaean classification level. I used 100%, 90%, and 80% for each cutoff in all possible combinations. Some examples of such classifications from Lake Mendota are shown in Table 1.

Example classification for a set of GFMs

GFM Name	domain	phylum	class	order	family	genus	species
MEint.metabat.1091	Bacteria	Actinobacteria	Actinobacteria	Acidimicrobiales	Acidimicrobiaceae	Ilumatobacter	
MEint.metabat.2538	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	Candidatus Aquiluna	
MEint.metabat.112	Bacteria	Cyanobacteria					
MEint.metabat.7672	Bacteria	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	Rhodopirellula	
MEint.metabat.3080	Bacteria	Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	Methylotenera	Methylotenera Versatilis
TBepi.metabat.3838	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Polynucleobacter	Polynucleobacter Necessarius

*Table 1:* Each GFM was classified by Phylosift (Darling et al., 2014) and custom script. Shown above is a subset of GFMs from Lake Mendota. The cyanobacterial GFM (MEint.metabat.112) is an example of one that does not follow Linnaean classification, and thus does not get labeled past phylum. The others stop getting classified when the next level does not match above the matching threshold specified.

Most genomes are not classified to the next level because they did not match above the matching percentage cutoff. However, some genomes, such as those classified to the cyanobacteria phylum, do not proceed to the more refined levels because their classification in the NCBI database does not follow the Linnaean structure. They are missing several levels and contain some levels of “no rank”. This is a downside to my program since it requires each Linnaean level and matching at that level to proceed to the next. However, this was necessary because otherwise one genome could have a classification where, for example, the genus did not belong to the order. For organisms of interest that were not sufficiently classified by this system, we can always go back and manually find the classification. Table 2 is a summary of how many genomes were classified to phylum and class and the number of phyla and classes represented from each lake/layer.

Number of GFMs classified

	Mendota	TroutBog_epi	TroutBog_hypo
# of genome bins	102	36	64
classified at phylum level	98	36	60
phyla represented	9	6	8
classified at class level	78	35	56
classes represented	14	11	15

Table 2: The first row in this table gives the number of genomes binned, after filtering based on completeness and uniqueness of single copy genes. The second and forth rows show the number of GFMs classified to the phylum and class level, respectively. The third and fifth rows give the number of phyla and classes represented in the GFMs by each lake/layer.

GFM phyla distribution between lakes/layers

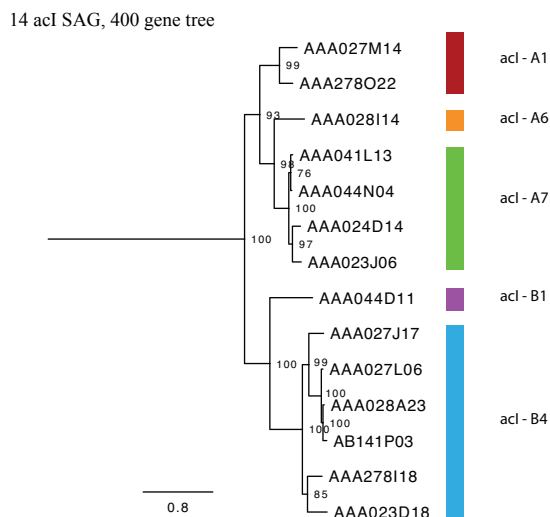
Phylum	Mendota	Trout Bog Epi	Trout Bog Hypo
ACIDOBACTERIA	0	2	3
ACTINOBACTERIA	17	9	9
BACTEROIDETES	33	3	9
CHLAMYDIAE	1	0	0
CHLOROBII	0	2	2
CHLOROFLEXI	1	0	0
CYANOBACTERIA	11	0	0
ELUSIMICROBIA	0	0	1
IGNAVIBACTERIA	0	0	2
PLANCTOMYCETES	13	0	0
PROTEOBACTERIA	12	17	26
ALPHAPROTEOBACTERIA	2	3	4
BETAPROTEOBACTERIA	7	9	11
DELTAPROTEOBACTERIA	1	1	4
EPSILONPROTEOBACTERIA	0	0	1
GAMMAPROTEOBACTERIA	2	4	4
TENERICUTES	2	0	0
VERRUCOMICROBIA	8	3	8

Table 3: The table above contains the distributions by phyla for the GFMs from each lake/layer. There are several phyla with only a few in one lake/layer and none in the others. However, cyanobacteria and planctomycetes have many GFMs represented in Lake Mendota and none in Trout Bog.

Shown in Table 3 is the breakdown of the number of genomes classified for a phylum by lake and layer. Though there are a number of phyla only represented in one of the two lakes, most of them are very low in numbers and could be different due to difference in assembly between the lakes. There are also a number of phyla only represented in the hypolimnion of Trout Bog, but these too could be different due to assembly. The two phyla that have several genomes represented from Lake Mendota and none from Trout Bog include cyanobacteria and planctomycetes. The presence of cyanobacteria in Lake Mendota and not in Trout Bog is unsurprising since the latter is a dystrophic lake and therefore is darkly stained and allows for little light penetration. This is also unsurprising since Lake Mendota is eutrophic

and is known to have cyanobacterial blooms throughout the summer (Beversdorf *et al.*, 2013). Also, from a preliminary analysis of our unpublished 16S data for both of these lakes, cyanobacteria have 6x the relative abundance in Lake Mendota than in Trout Bog. However, planctomycetes make up relatively the same portion of reads in the 16S data for both Lake Mendota and Trout Bog. Possible explanations include that planctomycetes did not assemble from Trout Bog due to differences in diversity, or that the differential coverage signal was not strong enough to bin GFMs. Both of these explanations highlight that these GFMs may not capture all the populations in our lakes.

### *Acl evolution*



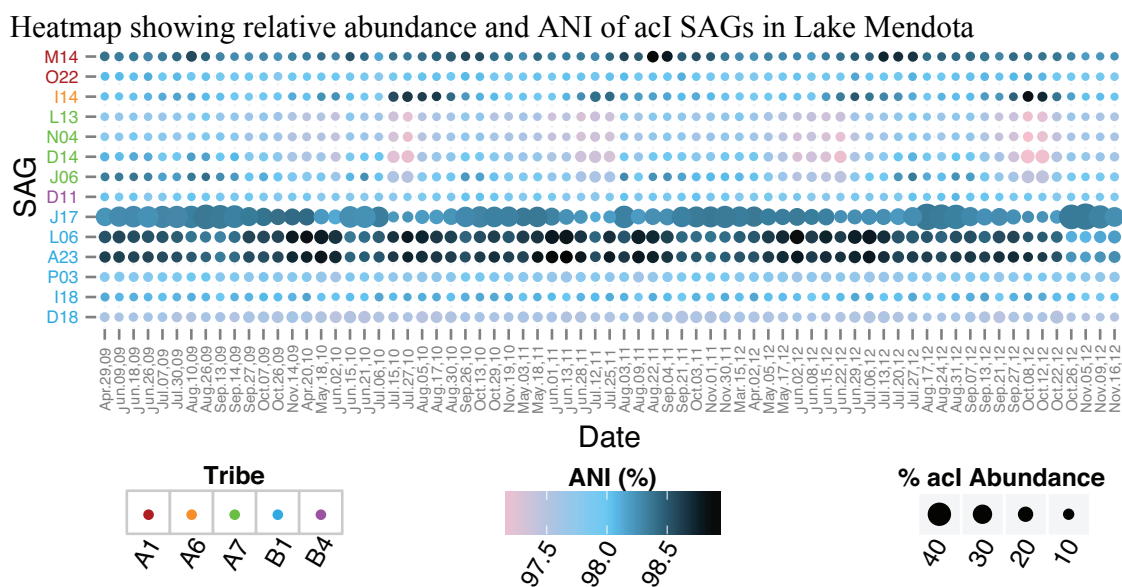
*Figure 1:* Using the alignment of 400 marker genes from PhyloPhlan (Segata, Børnigen, Morgan, & Huttenhower, 2013), a maximum likelihood tree was made with RAxML (Stamatakis, 2014). For this tree, 100 bootstrap replicates were run. The branching order for this tree, matches that of the 16S tree.

In collaboration with Dr. Sarahi Garcia, a postdoc in the McMahon lab, I have been working on an evolutionary analysis of SAGs from the *acl* lineage of actinobacteria. The *acl* lineage is a ubiquitous and often abundant member of freshwater lake communities (Garcia *et al.*, 2013). We analyzed 14 genomes from this lineage, which represent 5 of the 13 tribes as defined by 97% 16S similarity. Figure 1 shows a tree I made with RAxML (Stamatakis, 2014) from an

alignment of 400 conserved single copy genes made by PhyloPhlan (Segata *et al.*, 2013), which also resolves the genomes into the tribes previously defined by 16S similarity.

I first used BLASTP and MCL to cluster orthologous genes and we hoped to use this strategy to determine the core genome of the *acl* lineage and the core for some of the tribes within this lineage. However, we encountered a problem determining which genes were shared among all members of the lineage since the SAGs have varying levels of completeness. A current project, which developed out of this issue, is to determine if each cluster of orthologous genes is core for a group using the completion

estimates for the genomes in the group. I am currently working on this project in collaboration with PhD student Ben Oyserman in the McMahon lab. For the core determination and genome completion project I have developed a script to randomly delete portions of the genome to simulate the missing portions of a SAG. This program takes a complete genome, the percentage to delete, and a block size to delete. It then deletes that block size when possible or the number of bases left to be deleted when there is less than a block left to delete. It will finish when it has deleted the number of bases that make up the percentage you entered. Next, we hope to test this method on a simulated set of SAGs and a test set of SAGs from JGI.



**Figure 2:** Relative abundance and ANI for each SAG in each metagenome. Reads were mapped using BLAST. Only hits longer than 200 bp and higher than 95% ID were kept.

To continue on the *aci* project, Dr. Garcia and I decided to look at tracking these populations in nature by mapping our metagenomic reads back to the SAGs. I wrote a script to automate the process of running BLAST for each SAG against all of the metagenomes. We also proved to ourselves how the 95% average nucleotide identity(ANI) cutoff would track each SAG by shredding the SAGs and using BLAST to map these simulated metagenomes back to the most complete SAG from each tribe. Once we proved the cutoff, I filtered out BLAST hits below the 95% ANI limit from the recruitment of metagenome reads to SAGs. I then tracked relative abundance across the strains (Figure 2). I developed a script to parse the BLAST results and calculate the ANI of all the hits and the relative abundance. I also calculated the 16S

similarity and wrote a program to find average amino acid identity(AAI) for comparing the genomes to one another, see Figure 3. This experience will help me track the populations of other SAGs and GFMs using the metagenomic reads.

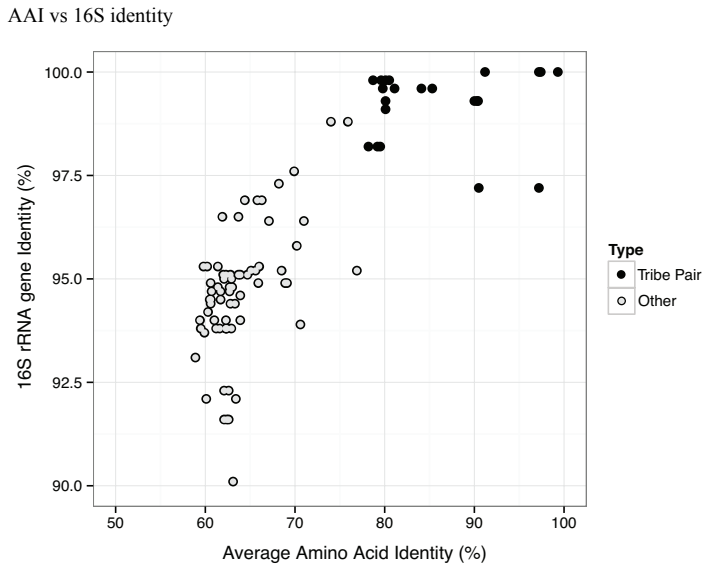


Figure 4: For every pair of acI SAGs, the AAI is plotted versus the 16S rRNA gene identity. Black indicates that both SAGs were from the same tribe. The 16S identity is not always a good predictor of AAI.

## Experimental Plan

### Aim 1:

**To determine the evolutionary dynamics of sequence discrete populations of bacteria in freshwater lakes, we will map reads from our metagenomic time series to composite GFMs and SAGs and examine abundance patterns, SNP patterns, gene gain and loss, and recombination.**

I will start by identifying GFMs and SAGs (henceforth referred to as genomes when both are used) with interesting abundance patterns based on the metagenomic mapping already done by our collaborators at JGI. Reads were mapped to each genome at 95% sequence identity, as this has previously been shown to correspond to sequence discrete populations when mapping metagenomic reads to reference strains (Caro-Quintero & Konstantinidis, 2012; Chan *et al.*, *in review*). However, I will test a subset of GFMs and all SAGs to check that this theory generally holds true across this dataset. The term population will be used to mean sequence discrete population for the remainder of this paper. Relative abundance for all the populations will be calculated using the coverage normalized by the size of the metagenome and size of the genome. The relative abundance inferred from the mapping is crucial for identifying if the bacterial populations are **persistent** or **transient** and **abundant** or **rare** in the lake. A

Mock Plot of Mean vs. CV

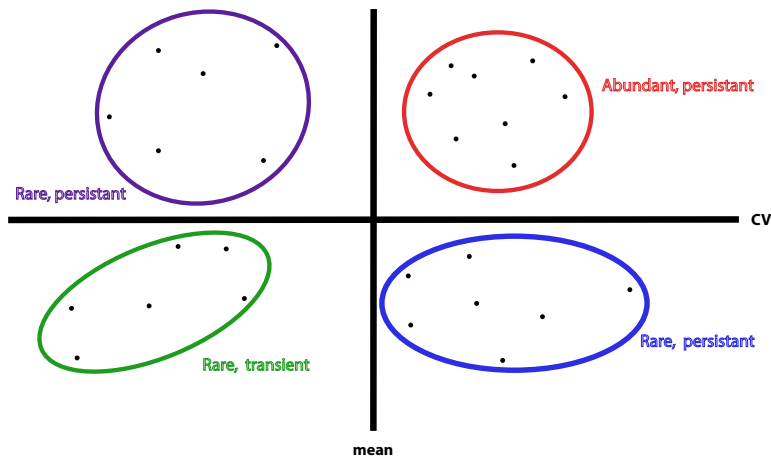


Figure 5: With mean relative abundance for each genome plotted against coefficient of variation of relative abundance, we should see a trend for which samples are abundant or rare and persistent or transient.

population will be categorized as **abundant** or **rare** by its mean relative abundance across all time points. A population of bacteria will be categorized as **persistent** or **transient** by the coefficient of variation (CV), which is standard deviation of relative abundance across all time points standardized by mean. The term **persistent** will mean a population

found be constant in abundance across time with lower CV, where the term **transient** will be applied to populations with large variation in abundance across time with higher CV. Once I have calculated both the mean and coefficient of variation (CV) for each population, I will be able to assign cutoffs relevant to this dataset. An example plot of mean verses CV is shown in Figure 5 and the quadrants are labeled **persistent** or **transient** and **abundant** or **rare**.

I will also use statistical inference to see if any of the abundance profiles vary predictably by season, biogeochemical conditions of the lake or relative abundance of another population. One might expect that the GFM's that are transient have some dependence that allows for blooms or causes the decline in population. Some possible causes for such blooms and busts include changes in biogeochemical conditions, changes in another population's relative abundance or an increase in predation such as by a population specific phage. With our dataset, we are unable to test the predation aspect. However, we can test dependence on biogeochemical conditions or the relative abundance of another population. The expectation of dependence is reversed for a persistent community member, which should tolerate a variety of lake conditions.

Since the two lakes in question are very different in their overall composition, one would also expect that the variation in average GFM relative abundance over time would be different for these two lakes. In fact, preliminary abundance profiles for the GFM's from the two lakes have different patterns

when coverage cutoffs are applied across samples. GFMs from Mendota generally have fewer time points that meet the coverage cutoff. A possible explanation for the variation in relative abundances for the GFMs is that a more biogeochemical stable lake has a more stable composition of GFMs (Hypothesis 1.1).

We can also search for genomes with interesting SNP patterns, as these have already been identified and quantified across time by our collaborators at JGI. I intend to look at the difference in numbers of SNPs between populations and their changes over time. Some possible patterns of intrapopulation variation, based on evolutionary models, are gene-specific or genome-wide sweeps. Sweep generally is defined as when all the SNPs in a population, either in a gene or whole genome, become the same variant over time. Populations experiencing high recombination rates compared to selective pressure would show gene-specific sweeps, whereas an advantageous gene variant sweeps the population. This is opposed to the ecotype model where genome wide sweeps occur with a higher selective pressure and lower rates of recombination since recombination cannot act to spread an advantage through the population in time. I will look for gene-specific or genome-wide sweeps in the persistent populations, as these metrics do not apply well to transient populations.

For the transient populations, we can look to see if different or the same variants are dominant for each bloom. I will look for this by examining the SNP patterns for transient populations. I expect to find cyanobacteria in the transient group of genomes from Lake Mendota, as we know the cyanobacteria undergo blooms throughout the summer (Beverdort *et al.*, 2013). I will be interested to see how the different genera of cyanobacteria bloom over the summer. I will search for the phycocyanin gene in these GFMs since it can be used to refine classifications for cyanobacteria (Miller *et al.*, 2013). I will also look for toxin genes which are known to be found in particular genera of cyanobacteria.

In order to look at the gene gain and loss patterns over time, the genes must be called and annotated for all the genomes. The GFMs are currently submitted for annotation in the JGI pipeline and are awaiting annotation. With annotations in hand, I can look for gene gain and loss over time using the metagenomes mapped to the genomes at 95% sequence ID, using methods similar to those used by our collaborators at JGI in their work with the manually binned genomes (Chan *et al.*, *in review*). This method normalizes the coverage by the gene length and excludes coding regions shorter than 450 bp. To find

the copy number of each gene per cell, I will divide the coverage of each gene by the average coverage for all the other genes in the genome, as was done previously by our collaborators at JGI. Genes will be considered lost from the genome if they rise above or below a certain threshold over all the time periods and are considered significant by the Metastats software (White *et al.*, 2009)

I also intend to look at recombination for these genomes in the environment. I will use a method for looking at the recombination of these populations, which also uses the metagenomic mapping. I will use this method to test the levels of recombination within these populations. Newton *et al.* showed that the lineage *acl* has much larger diversity in lakes than the tribe LD12. Due to this difference, I expect that *acl* will show higher rates of recombination and gene-specific sweeps whereas the LD12 tribe will show genome-wide sweeps due to lower rates for recombination, which was found previously seen in Zaremba-Niedzwiedzka *et al.* (Hypothesis 1.2, Hypothesis 1.3). I also expect that this trend will apply to the other populations studied (Hypothesis 1.4).

## **Aim 2:**

**To generate hypotheses about which bacterial groups are performing specific metabolic functions in the community, we will characterize the functional potential of bacteria using GFMs and SAGs.**

To make searching for metabolic potential in such a large number of genomes feasible, we need to develop a systematic way of looking at these genomes. Shaomei He, a scientist in the McMahon lab, is currently developing a scheme to do this. The method will be based on the system JGI's IMG uses to assess functional pathways and infer phenotypes (Chen *et al.*, 2013). The process will involve looking at multiple annotations for functions and finding a consensus among the systems.

Once a process is developed to methodically classify metabolic potential in genomes, I can then apply this system to the GFMs and SAGs. I plan on first identifying metabolic potential for genomes in aim 1 that showed interesting gene gain or loss patterns. I will also focus on specific phyla of interest to compare the different GFMs and see if there is any pattern among the larger evolutionary group.

I also plan to use a transporter specific process to identify uptake potential. I will use the TCDB database (Saier *et al.*, 2014) to find transporters for the substrates tested with MAR-FISH (Salcher *et al.*, 2013). The BioV collection of scripts was developed to work with the TCDB database. I will find



homologs to known transporters in the GFMs and SAGs from groups studied by MAR-FISH using the BioV collection. I will first specifically search for transporters that could perform the substrate uptake found by MAR-FISH. I can also use this method to determine if all bacteria in the groups tested take up the substrate or just a subset. I will also use the TCDB database to search for homologs to known transporters in GFMs and SAGs not tested in previous studies. These predictions can be used in conjunction with the metabolic potential predictions as a starting point for designing further MAR-FISH experiments.

Once the metabolic functions and transporters are predicted for the GFMs and SAGs, we can then look for functions specific to one group. Functions predicted to be found in specified groups will be the basis for MAR-FISH experiments. Alex Linz, a MDTP student in the McMahon lab, is currently working on learning MAR-FISH and preparing the lab to run these types of experiments.

### **Aim 3:**

**To determine if genome streamlining is a general characteristic of abundant freshwater bacteria, and if it is predicted by certain genomic features that I will search for in our GFMs and SAGs.**

Genomes will be considered more streamlined if they are both smaller and have a smaller proportion of intergenic spacer DNA to coding DNA. Each GFM and SAG will be evaluated by these criteria. Based on previous characterization of lake genomes, we expect certain genomic features may separate streamlined genomes from non-streamlined genomes (Livermore *et al.*, 2013). I will characterize these factors for the GFMs and the SAGs. The three features previously found to differentiate bacterial lifestyles in lakes are diversity of carbon substrate usage, growth rate as predicted by codon bias and proportion of signal transduction and motility genes in the genome.

With regards to carbon substrate utilization, I will use the carbon usage predicted for these genomes in aim 2. I will then cluster the results to see if the pattern of carbon substrate utilization groups by genome streamlining. I expect that the diversity of carbon substrate utilization will decrease as genome streamlining increases (Hypothesis 3.1). This can also be done with the other functions predicted in aim 2 to see if any other functions correlate with streamlining.

The second dimension that I will characterize is predicted growth rate. Since codon bias was found to be an accurate single predictor of growth rate (Vieira-Silva & Rocha, 2010), I will use it to calculate

predicted growth rate. I will create a program to calculate codon bias and growth rate when given a genome with its genes called. I started to create a similar program to calculate codon bias during my rotation in the Vetsigian lab. I can add the additional function of growth rate estimation using the method developed by Vieira-Silva et al., 2010. I will plot both streamlining values against growth rate to see how these relate. I expect that the growth rate will negatively correlate with streamlining (Hypothesis 3.2); as the genome becomes more streamlined the growth rate decreases.

The final previously characterized dimension that I will search for in these genomes is the proportion of signal transduction and motility genes for each of these genomes. I will use signal transduction and motility groups from the Cluster of Orthologous Genes (Tatusov *et al.*, 2000) assignments within IMG. I will normalize by the total number of genes with COG assignments. I plan to plot these values against the genome streamlining values to see how these correlate, if at all. I expect that genome streamlining will correlate negatively with proportion of genes assigned to the motility and signal transduction COG; as streamlining increases, the proportion of genes for motility and signal transduction decreases (Hypothesis 3.3).

Based on previous calculations of effective genome size using metagenomes, I expect that genome streamlining is common among abundant freshwater bacteria (Hypothesis 3.4). In previous calculations of effective genome size, both Lake Mendota and Trout Bog, had effective genome sizes of less than 3 megabases (Eiler *et al.*, 2013). To test this hypothesis, I will see if the abundance categorizations from aim 1 correlate with genome streamlining.

## Timetable

Year	2nd	3rd	3rd	3rd	4th	4th	4th	5th	5th
Season	Su	Fa	Sp	Su	Fa	Sp	Su	Fa	Sp
AIM 1									
Rel. Abund.									
SNP patterns									
Gene gain/loss									
Recombination									
AIM2									
Metabolic Potential									
AIM 3									
Carbon Diversity									
Growth Rate									
Sign. Trans. + Mot.									
Other features									
Paper Writing									
Mentoring									
Conferences									

## Literature Cited

- Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**. <http://www.nature.com/doi/10.1038/nbt.2579> (Accessed May 26, 2013).
- Amann RI, Ludwig W, Schleifer KH, Amann RI, Ludwig W. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation . Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation. *Microbiol Mol Biol Rev* **59**.
- Beverdort LJ, Miller TR, McMahon KD. (2013). The role of nitrogen fixation in cyanobacterial bloom toxicity in a temperate, eutrophic lake. *PLoS One* **8**:e56103. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3566065&tool=pmcentrez&rendertype=abstract> (Accessed May 6, 2014).
- Caro-Quintero A, Konstantinidis KT. (2012). Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**:347–55. <http://www.ncbi.nlm.nih.gov/pubmed/22151572> (Accessed February 4, 2014).
- Chan L-K, Bendall ML, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, *et al*. Genome-wide selective sweeps in natural bacterial populations revealed by time-series metagenomics. *ISME J*.
- Chen I-M a, Markowitz VM, Chu K, Anderson I, Mavromatis K, Kyrpides NC, *et al*. (2013). Improving microbial genome annotations in an integrated database context. *PLoS One* **8**:e54859. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3570495&tool=pmcentrez&rendertype=abstract> (Accessed March 9, 2013).
- Cohan FM, Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**:R373–86. <http://www.ncbi.nlm.nih.gov/pubmed/17502094> (Accessed February 20, 2014).
- Cordero OX, Polz MF. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12**:263–73. <http://www.ncbi.nlm.nih.gov/pubmed/24590245> (Accessed April 28, 2014).
- Darling AE, Jospin G, Lowe E, Matsen F a., Bik HM, Eisen J a. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**:e243. <https://peerj.com/articles/243> (Accessed January 21, 2014).
- Eiler A, Zaremba-Niedzwiedzka K, Garcia MM, McMahon KD, Stepanauskas R, Andersson SGE, *et al*. (2013). Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. *Environ Microbiol*. <http://www.ncbi.nlm.nih.gov/pubmed/24118837> (Accessed April 28, 2014).
- Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas R, *et al*. (2013). Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *ISME J* **7**:137–47. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3526179&tool=pmcentrez&rendertype=abstract> (Accessed March 19, 2013).
- Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, *et al*. Comparative single-cell genomics reveals potential ecological niches for the freshwater actinobacteria lineage.
- Gilbert J a, Dupont CL. (2011). Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci* **3**:347–71. <http://www.ncbi.nlm.nih.gov/pubmed/21329209> (Accessed April 28, 2014).

- Giovannoni SJ, Cameron Thrash J, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* 1–13. <http://www.ncbi.nlm.nih.gov/pubmed/24739623> (Accessed April 28, 2014).
- Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**:1242–5. <http://www.ncbi.nlm.nih.gov/pubmed/16109880> (Accessed February 28, 2013).
- Grote J, Thrash J, Huggett M, Landry Z. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* **3**:1–13. <http://mbio.asm.org/content/3/5/e00252-12.short> (Accessed April 28, 2014).
- Livermore J a, Emrich SJ, Tan J, Jones SE. (2013). Freshwater bacterial lifestyles inferred from comparative genomics. *Environ Microbiol.* <http://www.ncbi.nlm.nih.gov/pubmed/23889754> (Accessed November 4, 2013).
- Miller TR, Beversdorf L, Chaston SD, McMahon KD. (2013). Spatiotemporal molecular analysis of cyanobacteria blooms reveals microcystis--aphanizomenon interactions. *PLoS One* **8**:e74933. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3785500&tool=pmcentrez&rendertype=abstract> (Accessed May 12, 2014).
- Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE, *et al.* (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**:784–98. <http://www.ncbi.nlm.nih.gov/pubmed/20072162> (Accessed March 31, 2014).
- Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the natural history of freshwater lake bacteria. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3063352&tool=pmcentrez&rendertype=abstract> (Accessed March 6, 2013).
- Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**:431–7. <http://www.ncbi.nlm.nih.gov/pubmed/23851394> (Accessed March 19, 2014).
- Saier MH, Reddy VS, Tamang DG, Västermark A. (2014). The transporter classification database. *Nucleic Acids Res* **42**:D251–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3964967&tool=pmcentrez&rendertype=abstract> (Accessed April 30, 2014).
- Salcher MM, Posch T, Pernthaler J. (2013). In situ substrate preferences of abundant bacterioplankton populations in a prealpine freshwater lake. *ISME J* **7**:896–907. <http://www.ncbi.nlm.nih.gov/pubmed/23235289> (Accessed November 4, 2013).
- Segata N, Börnigen D, Morgan XC, Huttenhower C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**:2304. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3760377&tool=pmcentrez&rendertype=abstract> (Accessed May 8, 2014).
- Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–3. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3998144&tool=pmcentrez&rendertype=abstract>.

Tatusov RL, Galperin MY, Natale D a, Koonin E V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**:33–6.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102395&tool=pmcentrez&rendertype=abstract>.

Vieira-Silva S, Rocha EPC. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* **6**:e1000808.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2797632&tool=pmcentrez&rendertype=abstract> (Accessed November 4, 2013).

Viklund J, Ettema TJG, Andersson SGE. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* **29**:599–615.  
<http://www.ncbi.nlm.nih.gov/pubmed/21900598> (Accessed March 19, 2013).

White JR, Nagarajan N, Pop M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* **5**:e1000352.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2661018&tool=pmcentrez&rendertype=abstract> (Accessed April 28, 2014).

Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**:e5299.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2668756&tool=pmcentrez&rendertype=abstract> (Accessed March 21, 2014).

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, *et al.* (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**:1661–5. <http://www.ncbi.nlm.nih.gov/pubmed/23019650> (Accessed February 28, 2013).

Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, *et al.* (2013). Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol* **14**:R130. <http://www.ncbi.nlm.nih.gov/pubmed/24286338>.