**Abstract**

**Specific Aims**

Goal: Learning about the ecology and evolution of microbes in lakes using genomics, including how populations change over time in lakes, metabolic potential of the lake microbes, and genomic features of the lake microbes.

**AIM 1:**

Bacterial evolution in nature has not previously been well studied because many 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 acI lineage of Actinobacteria. Using our metagenomic time series, which spans multiple years in two very different lakes (Trout Bog and Mendota), we can use coverage based binning and sequence features to retrieve reference genomes, which are directly relevant to our data set and not necessarily from cultivatable lineages. We also have single amplified genomes (SAGs), which represent 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. This approach is novel because coverage based binning techniques have only recently been developed. The 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.

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

*H1.1 Lakes with more stable biogeochemical conditions over time, also have less variation in relative abundance for the all the GFMs from that site.*

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

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

*H1.4: Each lineage 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 previous been involved in looking at functional capabilities of the acI-Actinobacteria lineage using SAGs(Sarahi’s paper, Trevor’s Paper). We can also search for temporal correlations between nutrient conditions in the lakes and relative abundances of 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.

**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 cultivatable. 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 cultivatable bacteria may not be representative of those microbes that are most abundant or are performing important functions in the environment. 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 give a better understanding how important genome reduction is to be successful in aquatic environments. This work will give a better understanding of genomic streamlining as a microbial lifestyle and how it functions in a community setting.

**To learn about genomic streamlining and its prevalence among abundant lake bacteria, we will characterize the features of genome streamlining in uncultivated bacterial genomes.**

*H3.1: Genome size correlates negatively with growth rate.*

*H3.2: As genome size decreases so do the genes allowing for a diversity of carbon substrate utilization.*

*H3.3: As genome size decreases so do the genes associated with motility or signal transduction.*

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

**Background and Significance**

Methods for studying the uncultivated majority of microbes -

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, Ludwig, Schleifer, Amann, & Ludwig, 1995). The development of molecular techniques, such as high throughput sequencing and genomics, has enabled us to start to studying which microbes reside in an environment are and what they are doing. Many approaches have allowed scientists to peek into the uncultivated world, such as 16S tag sequencing, metagenomics, and sequencing of single cells. Each method has different advantages and disadvantages, but can be used to answer specific questions about the microbes in an environment.

One method of learning which microbes are in a specific environment is sequencing a conserved marker such as the 16S rDNA or other conserved genes. In freshwater lakes, 16S 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 develop a 16S database(Newton, Jones, Eiler, McMahon, & Bertilsson, 2011). While this approach, with a good reference database, can help to answer the question of which microbes inhabit the lake, it does 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.

Another tool that can yield genomic information about the uncultured microbes is metagenomics. This method involves extracting DNA from lake water samples. It can provide a representation of the lake community not biased by cultivation. One limitation of this technique is that the data is difficult to track back the reads or even assembled contigs to individual organisms or closely related populations. Previous studies relied on sequence features such as GC content and tetranucleotide frequency to separate out contigs belonging to one lineage. A recently developed technique uses differential coverage to separate out contigs that belong to the same organism(Albertsen et al., 2013; Wrighton et al., 2012). This method requires that you have two or more different metagenomes representing the same environment. Possible differences include the extraction method or time of sample collected.

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). These contigs were binned into genomes first by emergent self-organizing maps of tetranucleotide frequency and then by differential coverage. From these genomes, metabolic potential was inferred from annotated genes.

Another such study used the differential coverage as the first metric for binning genomes from an activated sludge bioreactor(Albertsen et al., 2013). Then the bins were refined using tetranucleotide frequency, GC content, length, and single copy gene analysis. When this method was applied to the dataset from the aquifer, they found that this method assembled more complete genomes with less contamination from other genomes. 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 computation. This method was applied to produce 201 SAGs from previously unsequenced lineages (Rinke et al., 2013). These phyla had been known to exist in nature due to 16S tag sequencing but have never been isolated. These genomes provide diversity and expand the reference database to include these phyla. In addition, 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). These SAGs recruited reads much better than cultured marine flavobacteria. The authors mapped 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 important example of how it is important 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 (Albertsen et al., 2013; Garcia et al., 2013; Rinke et al., 2013)+TEVOR’s PAPER.

Studying evolution using genomics –

This rise of sequencing has given us the opportunity to study evolution by genome comparison. From this data single mutations can been found, and difference in functional ability inferred. With cultured representatives, *in situ* tests can be done to test fitness (cite lenski experiments?). One example of studying evolution in the laboratory is a long term experiment transferring *E. coli*. Serial transfers provide population bottlenecks regularly. EXPERIMENTS AND SEQUENCING?- find paper and look for info.

Previous studies have also studied evolution by isolating many genomes from those environments and comparing their mutation and recombination. VIBRIO AND SULFOLOBUS EXPERIMENTS – see Rex’s references

Attempts have also been made using metagenomics to follow populations in nature and calculate their mutation rates. ACID MINE METAGENOMICS – SEE REX’s references

A new study done by the Joint Genome Institute in collaboration with the McMahon lab, provides evidence for the ‘ecotype’ model of bacterial speciation(Reference REX PAPER). This work tracks single nucleotide polymorphisms(SNPs) over three years, in four genomes binned from metagenomes by mapping the reads of these metagenomes back to the genomes. Reads were mapped at ninety-five percent identity representing sequence discrete populations in the lake. Over the three-year time period all the SNPs in three of the populations go to fixation. Thus the sequence discrete populations represent ecotypes swept by the same selective pressure.

Another current study in the McMahon lab using genomics to study evolution focuses on the Actinobacterial clade acI, a ubiquitous and abundant lake bacterium. This study uses 14 SAGs from the acI clade and tracks them over time by mapping back reads from a 5 year time series of Lake Mendota. To understand what we are seeing when reads are mapped backed to these genomes from the metagenomes, we mapped shredded reads from the SAGs and mapped them back to one SAG from each tribe. DO I NEED TO EXPLAIN TRIBE? These shredded SAGs together represent a pseudo-metagenome. The mapping results then show if reads from different or the same tribes are recruited by BLAST. We found that only

Studying ecology using genomics –

**Preliminary Results**

*Genomes from Metagenomes*

While at JGI, I worked with collaborators, developing a program called Metabat to bin genomes from metagenomes. This program uses differential coverage across the XX metagenomes for the epilimnion of Trout Bog Lake, the YY metagenomes from the hypolimnion of Trout Bog Lake, and the ZZ metagenomes 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 bounded by the program and require no manual intervention to select bins as was done with other binning methods. From this program I have 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 90% of those single copy genes were unique. After this filtering, there are 36, 70, and 104 GFMs in Trout Bog epilimnion, Trout Bog hypolimnion, and Lake Mendota, respectively.

In order to classify these genomes, I used a program called Phylosift(CITATION). It finds 37 conserved evolutionary marker genes using hidden markov models, aligns them, and gives a probability of classification for each maker gene. It is designed to be used on metagenomes and give you an idea of which organisms are in your 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 have matching results above the percent matching threshold, it will save that classification and proceed to the next Linnaean classification level. I used a range of values of 100%, 90%, and 80% for each cutoff. Some examples of such classifications from Lake Mendota are shown in Table 1.

Most genomes are not classified to the next level because it did not match at the matching percentage cutoff. However, some genomes, such as the one classified to the Cyanobacteria phylum, don’t proceed to the later levels because their classification in the NCBI database does not follow the Linnaean classification. It is missing several levels and contains 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 didn’t belong the order. For organisms of interest that were not sufficiently classified by this system, one could 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 represented from these groups.

Also shown is the breakdown of the number of genomes classified for phylum and lake/layer in Table 3. 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 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 difference due to assembly. Two phyla that have several genomes represented from 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 little light penetration. This is also unsurprising since Lake Mendota is eutrophic and is known to have cyanobacterial blooms throughout the summer.

*AcI evolution*

In collaboration with a postdoc in the McMahon lab, Sarahi Garcia, I have been working on an evolutionary analysis of SAGs from the acI lineage of Actinobacteria. The acI lineage is common and often abundant member of freshwater lakes. We analyzed 14 genomes from this lineage which represent 6 of the 13 tribes as defined by 97% 16S similarity. Figure 1 shows a tree of 400 conserved single copy genes, which also resolves the genomes into the tribes previously defined by 16S similarity. We first used BLASTP and MCL to cluster orthologous genes. We hoped to use this strategy to determine the “core” genome of the acI lineage and some of the tribes within this lineage.

However, we encountered a problem when determining which genes were shared among all members of the lineage since the SAGs are varying levels of completeness. A current project developed out of this issue to statistically estimate how likely any cluster of orthologous genes is to be core given the completion estimates of the genomes missing a gene in that cluster. I am currently working on this project in collaboration with Ben Oyserman in the McMahon lab. For that project I have developed a script to randomly delete portions of the genome to simulate the missing portions of the genome. 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 that need to be deleted when it is no longer possible to delete blocks. It will finish when it has deleted the number of bases to get it to the percentage you wished to delete. Next, we hope to test this method on a simulated set of SAGs and a test set of SAGs from JGI.

To continue on acI, we 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 compete SAG from each tribe. Once we proved the cutoff, we used the 95% ANI limit for BLAST hits from metagenome reads to SAGs (Figure 2). We then tracked relative abundance across the strains (Figure 3). I developed a script to parse the BLAST results and calculate the ANI and 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 for tracking the populations of other SAGs and GFMs using the metagenomic reads.

**Experimental Plan**

**Aim 1:**

**To determine the evolutionary dynamics of sequence discrete populations of bacteria in freshwater lakes, we will map reads from 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 with interesting abundance patterns based on the metagenomic mapping already done by our collaborators at JGI. Reads will be mapped at 95% sequence identity as this has previously been shown to correspond to sequence discrete populations when mapping metagenomic reads to reference strains (CITE KOSTAS and REX paper). However, I will test a subset of GFMs to check that this generally holds true in this dataset. The relative abundance inferred from the mapping is crucial for identifying if the bacterial populations are persistent or transient in the lake. Transient bacteria are undergoing selection that reduces the relative amount of the lineage in the community. This is opposed to the abundance profile experience by the persistent bacteria that have insignificant changes in abundance across time. To find if a population is persistent or transient, I will quantify the average change in abundance over all time points for all GFMs from the same site and classify if each GFM is more or less persistent than average. I would expect that the contigs that make up a single GFM would all be either persistent or transient by this metric, since differential coverage was used in binning these genomes. However, I will investigate the validity of the binning should a contig not follow a similar pattern as the others in that same GFM. If any contigs with different persistence are determined to actually belong in that GFM, I will use the majority persistence for all contigs to classify the GFM.

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 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 a population specific phage. 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 might be different for these two lakes. In fact, preliminary abundance profiles for the GFMs 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 (H1.1).

We can also search for GFMs with interesting SNP patterns, as these have already been called and quantified across time by our collaborators at JGI. I intend to look at the difference in numbers of SNPs between populations and the changes over time. COME BACK HERE!!!

In order to look at the gene gain and loss patterns over time, the genes must be called an annotated for all the genomes. The GFMs are currently submitted for annotation in the JGI pipeline and are awaiting annotation. Once these genes are called, I can look for gene gain and loss over time by looking again using the metagenomes mapped at 95% sequence ID and methods similar to those used by Chan et al. with the manually binned genomes. 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. 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 (CITE METASTATS SOFTWARE, stole from Rex and weird article).

**Aim 2:**

**To hypothesize 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:**

**To learn about genomic streamlining and its prevalence among abundant lake bacteria, we will characterize the features of genome streamlining in uncultivated bacterial genomes.**

**Timetable**

**Literature Cited**