GEODES Project Status

# Sequencing

Samples collected in summer of 2016 were extracted and submitted for sequencing. 29 samples failed QC at JGI due to low yield, largely from Sparkling Lake and from the last timepoints of Trout Bog. These samples were replaced by replicates where available and where not, 4th replicates from timepoints with excellent yield were sent in their place. Samples for single cell sorting and sequencing (2 per lake) were sent in September. Samples for 16S sequencing (iTags) were sent in November and include samples from all three lakes all summer extracted with both the phenol-chloroform protocol and the FastDNA protocol.

# Bioinformatics

The major steps in the processing of the metatranscriptomes are:

* Quality filtering
* Removal of rRNA reads
* Mapping
* Counting of mapped reads

Quality filtering had already been performed by JGI – all subsequent steps were performed on JGI’s QC filtered fastq files. rRNA reads were removed using sortmerna, with all alignments accompanying the program installation as references. Mapping was performed using a custom, curated, non-redundant gene database (more to follow) using bbmap. Multi-mapping reads are counted, but randomly assigned. Mapped reads are counted using FeatureCounts, with no overlaps or fractional reads allowed. The resulting table is used as the input for analyses in R.

The non-redundant gene database was created using the metagenome assemblies generated as part of GEODES, the MAGs and SAGs generated by the McMahon Lab in previous projects, genomes of eukaryotic algae from the refseq database (as our existing genomes are predominantly bacterial), and the SAGs generated as part of GEODES (yet to be sequenced). The non-redundant gene database approach reduces both the size of the database and the proportion of multi-mapping reads, while still retaining genome-level information about each gene.

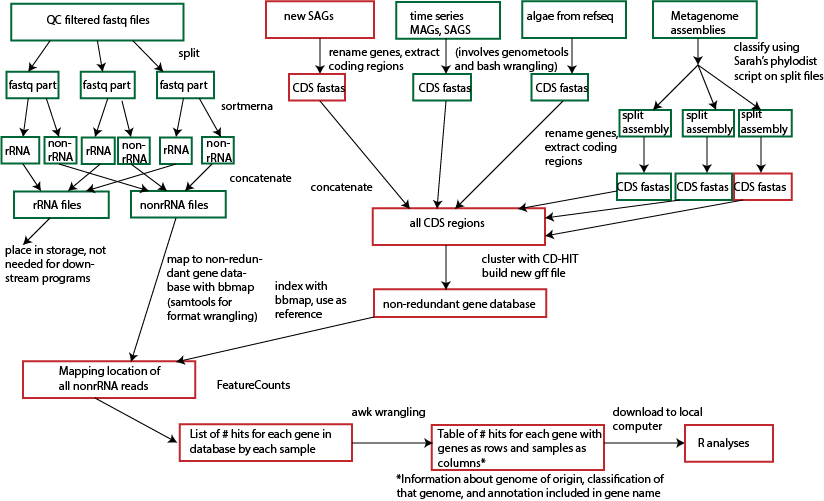
Before inclusion in the database, each type of genome must first have its GFF file formatted. GFF files were formatted, sorted, and standardized with genometools. During this step, the gene IDS were changed to include relevant information such as product names, taxonomic classification, and genome of origin, as this will be most easily searchable once mapped reads have been counted up by gene. Previous to the GFF formatting of metagenome assemblies, the contigs were reclassified based on USEARCH results of each gene on each contig against all genomes in the IMG database. The consensus of the classification of the best hits for contigs with at least three genes was assigned to the whole contig. Once genes were renamed and gff files reformatted, the corresponding regions were extracted from the fasta files using genometools, concatenated across all types of genomes to be included, and clustered using CD-HIT. A representative gene for each cluster is randomly chosen by CD-HIT and included in the mapping database. A new gff file must be manually created before the database can be used for mapping.

All bioinformatics analysis was performed through the UW-Madison Center for High Throughput Computing, which is based on the Condor system. Condor is well-suited for parallelizing many small tasks, so the bioinformatics programs described above were run in parallel whenever possible. Sometimes this involved running all the samples or genomes in parallel, but when the task required was too large to be completed on a full sample, the sample itself was split. For example, rRNA sorting and removal was performed on subsets of samples, as aligning each read and deciding “rRNA” or “not rRNA” does not depend on any other reads. This allowed me to distribute the task across 7,000 CPUs rather than 110, completing the task in about 2 hours. Conversely, tasks such as indexing the mapping database cannot be split; in this case, the task is run on 1 CPU over several days, and efforts were taken to reduce the size of the database and the memory requirements of bbmap.

All code and notes from the bioinformatics workflow can be found at <https://github.com/McMahonLab/geodes>

My working version is located here: <https://github.com/alexlinz/geodes>

The 16S rRNA ribosomal amplicon data (iTags) is currently being analyzed by Robin Rohwer as part of a larger 16S workflow on lake data.



# Analysis

The goal of this project is to use diel trends, both in genes and organisms, to identify interactions between freshwater bacteria, steps in the freshwater carbon cycle, and differences in carbon cycling between lakes of different trophic statuses. The two main approaches to analyze diel trends will be network analysis and rhythmicity analysis. Network analysis will performed using weighted gene correlation network analysis in the R package WGCNA. With each gene as a node in this network, nodes can be color-coded by phylogenetic group, allowing concurrent analysis of diel trends at the gene and organism levels. Rhythmicity analysis will be performed using the R package RAIN, which uses non-parametric tests.

The results of GEODES can also be used to support other projects in the lab. The table of mapped read counts will be made available to all lab members so that people can look up the expression levels of their favorite organisms or genes. SAGs and metagenome assemblies will also be available for genome analysis. While binning of the metagenome assemblies is not part of this project (and is a computationally intense task), it could likely be used to assemble genomes from specific phylogenetic groups based on my classifications of contigs.

The purpose of the iTags in GEODES is to confirm that the days of metatranscriptome sampling were not “abnormal” days in the lakes. This will likely be tested via Bray-Curtis dissimilarity or UniFrac, and will be either a statement or supplemental figure in the final manuscript.

# Environmental data

Many environmental variables were tested concurrent with RNA collection, as these could potentially be drivers or results of observed trends in expression. The available environmental data, its collection method, and its observed trends are listed below.

|  |  |  |
| --- | --- | --- |
| Data Type | Collection method | Trends |
| High resolution water column and meteorological data | Instrumented buoy | Not yet analyzed; can be used to look at days before and after GEODES sampling |
| Dissolved oxygen and temperature depth profiles | Sonde | Confirms appropriate sampling depth; slight diel trends, particularly in Mendota |
| pH and conductivity | Sonde | These two variables appear correlated, but trend varies by lake and depth |
| Photosynthetically active radiation (PAR) | PAR meter (could be supplemented by local airport PAR measurements) | Strong diel trends; some missing data due to equipment malfunction |
| Chlorophyll | Filtration + methanol extraction | Diel in Trout Bog but not the other lakes; average concentration varies by lake |
| Protein production | C14-leucine bacterial production assay | Large variation in production level, but not diel |
| Total and dissolved nitrogen and phosphorus | Colorimetric assay by LTER | Slight diel trend that needs to be confirmed; levels vary by lake |
| Cyanotoxin profile | Tested by Todd Miller’s lab via HPLC for Mendota only | Diel trend maybe observed in one toxin, but very slight; mostly useful to provide information about which cyanotoxins are most abundant |