Genomics&Bacteria

Omics Analysis

A profile of the pond was constructed over a day-night cycle on June 25-26, 2015 at depths 0.5m, 2.0m, 3.0m, 4.0m, 6.0m, 8.0m, 10.0m, and 12m. At each time point, water was filtered for metagenomics and metatranscriptomic work to determine allocation of molecular machinery to primary metabolic pathways, as well as compare model predictions to both biogeochemical and genomic observations over time and space in Siders Pond.

In order to capture the genomic content of the most abundant organisms in Siders Pond, we carried out deep metagenomic sequencing of the microbial community at all 8 depths. The metagenome was only characterized at the first time point in the series as we did not expect population structure to change significantly over the 24 hour period. DNA was extracted from the first time point for all depths. Metagenomic libraries were constructed using Nugen multiplex systems and paired-end sequencing was performed on an Illumina MiSeq at the W.M. Keck sequencing facility at the Marine Biological Laboratory.

From this initial time point, we also generated a metatranscriptome for each depth to directly compare genomic potential and gene expression patterns of microbial communities at each depth. These results are critical to developing the model. RNA was extracted and metatranscriptomic libraries were constructed using Nugen multiplex systems and paired-end sequencing was performed on an Illumina MiSeq at the W.M. Keck sequencing facility at the Marine Biological Laboratory.

After assessing the geochemical and physical variability between time points, we chose depths 0.5 m, 3m, 6m, 8m, and 12 m for metatranscriptomic sequence for all seven time points to examine how gene expression patterns change over time and space. In total, we have sequence 46 samples, resulting in 8 metagenomes and 38 metatranscriptomes.

	Cast1	Cast2	Cast3	Cast4	Cast5	Cast6	Cast7
srf	Cast1Dsrf	Cast2Dsrf	Cast3Dsrf	Cast4Dsrf	Cast5Dsrf	Cast6Dsrf	Cast7Dsrf
2m	Cast1D2m						
3m	Cast1D3m	Cast2D3m	Cast3D3m	Cast4D3m	Cast5D3m	Cast6D3m	Cast7D3m
4m	Cast1D4m						
6m	Cast1D6m	Cast2D6m	Cast3D6m	Cast4D6m	Cast5D6m	Cast6D6m	Cast7D6m
8m	Cast1D8m	Cast2D8m	Cast3D8m	Cast4D8m	Cast5D8m	Cast6D8m	Cast7D8m
10m	Cast1D10m						
12m	Cast1D12m	Cast2D12m	Cast3D12m	Cast4D12m	Cast5D12m	Cast6D12m	Cast7D12m
	DNA/RNA	RNA	RNA	RNA	RNA	RNA	RNA

For each sample library, quality trimmed sequences were assembled using IDBA-UD using the default parameters. Because this was a 1 year project, we are currently still analyzing the large amount of -omics data generated. Here we describe the analysis currently underway.

First, we are determining the taxonomic distribution of each metagenome and metatranscriptome by mapping the reads of each sample to the Silva SSU and LSU Parc databases, followed by mapping of reads that matched the Silva databases to the Greengenes 16S rRNA database. In both cases, we are using bowtie2 for mapping using default settings and local alignment. Reads that mapped to the Greengenes database are being classified using mothur using the classify.seqs command against the Silva 16S rRNA database, using a cutoff of 50.

Second, each of the samples is being processed through the IMG/M annotation pipeline (Project Study. IMG/M annotations are being used for all searches involving metabolisms of interest and assignments of phylogenetic markers. Data is currently located in the JGI IMG portal entitled "Aquatic microbial communities from different depth of meromictic Siders Pond, Falmouth, Massachusetts" under study Gs0116874. The data is not currently public- IMG releases annotations to the public after 2 year. The raw reads will be deposited into the NCBI Short Read Archive later this summer.

Third, we are mapping the reads of each sample to the assembled contigs of every other sample using bwa aln, version 0.5.5, using default settings for mapping. We are using anvi'o version 1.2.1 to cluster contigs of each sample into bins. Anvi'o is a metagenomics binning and visualization platform that clusters contigs into bins based on tetranucleotide abundance and relative coverage. For this study, we are assembling reads from each sample individually rather than combine samples in a co-assembly. However, the relative coverage of each contig across all samples will be taken into account in order to improve the clustering of contigs into bins for each sample. We found that extracting bins from individual samples yielded longer contigs and more bins than co-assemblies, most likely due to high levels of intra-strain diversity. Reducing diversity by examining samples separately improved assembly and binning outputs. Bins are being selected in a supervised fashion, using a threshold of <10% redundancy based on the presence of universal single-copy genes as defined by four separate HMM profile collections in anvi'o. We will not include bins that are less than 20% complete, as determined by the presence of universal single-copy gene sets.

This workflow will allow us to determine allocation of molecular machinery for primary metabolic pathways, as well as compare model predictions to both biogeochemical and genomic observations over time and space in Siders Pond.

Preliminary Metagenomics and Bacterial Counts Results

DAPI cell counts from Sider Pond Samples show bacterial abundances vary from 2 to 25 x 10^6 cells mL^{-1}, with the greatest abundance in the deepest (12 m) samples (Fig. GB-1). Metagenomic analysis from Cast 1 of the 16S gene shows significant vertical gradients in taxa (Fig. GB-2), with green sulfur bacteria (*Chlorobi*) dominating the bacterial population at 8 m, where anoxic and light levels favor their dominance (Fig. GB-3). Preliminary identification of circadian rhythm genes in the Cast 1 metagenomics library (Fig. GB-4) revels 5 genes that vary in relative abundance over depth, with the greatest abundance occurring around 4 m that corresponds to oxygen and Chl a maximum zones (see Figs. FO-5 and 7 in FieldObservations.pdf). Research this summer by Petra Byl will be focus on bioinformatics of the substantial genomic and metagenomics libraries to complement MEP modeling work.

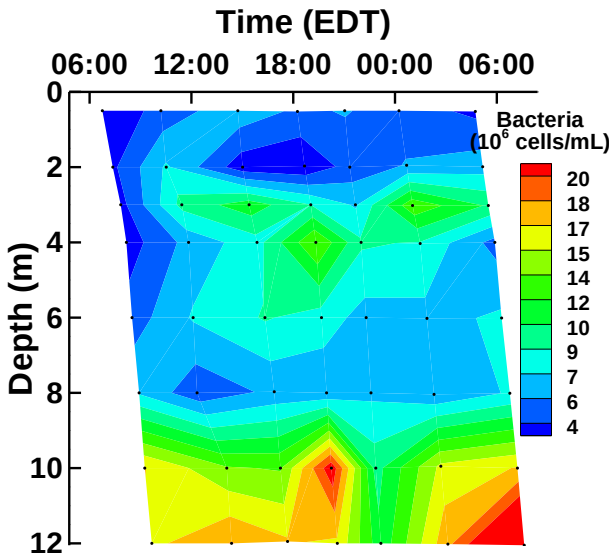

Fig. FB-1. Bacterial cell concentrations from Siders Pond field sampling in 2D plot.

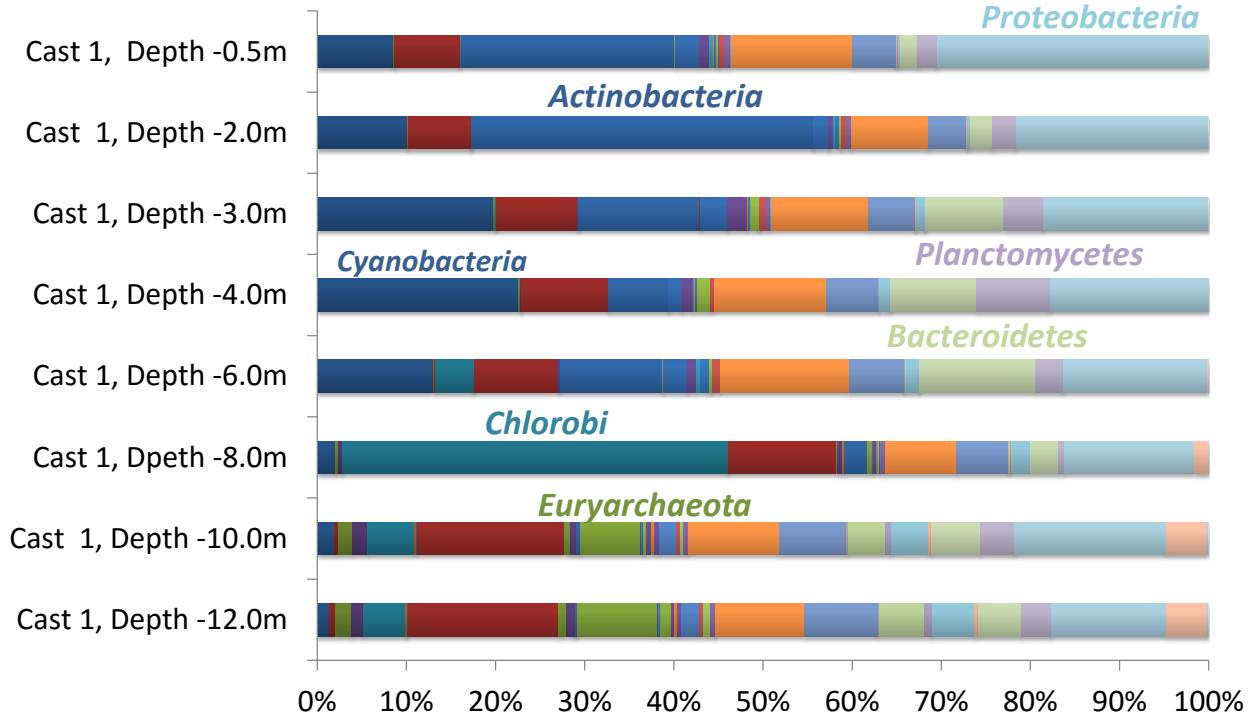

Fig. GB-2. Taxonomic profiles from metagenomics analysis of 16S genes from Cast 1 samples.

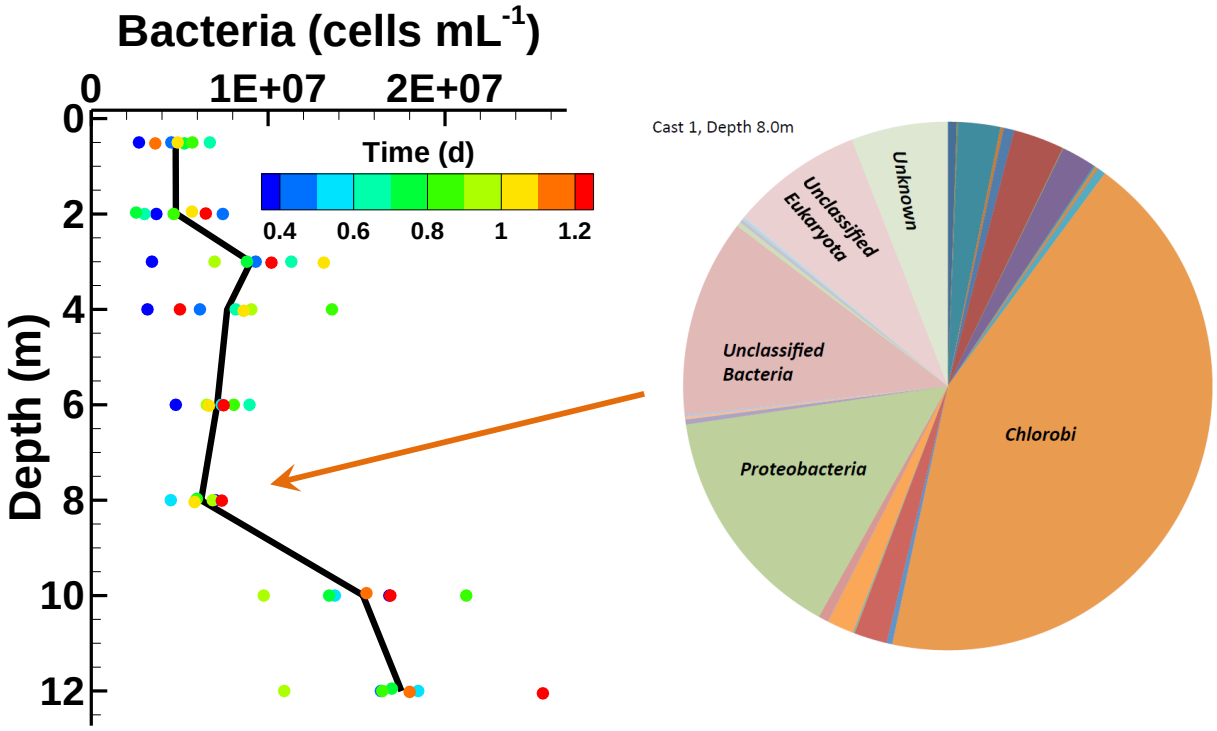

Fig. GB-3. Bacterial cell counts from all casts, and the taxonomy of the bacterial population at 8 m showing the dominance of green sulfur bacteria (*Chlorobi*).

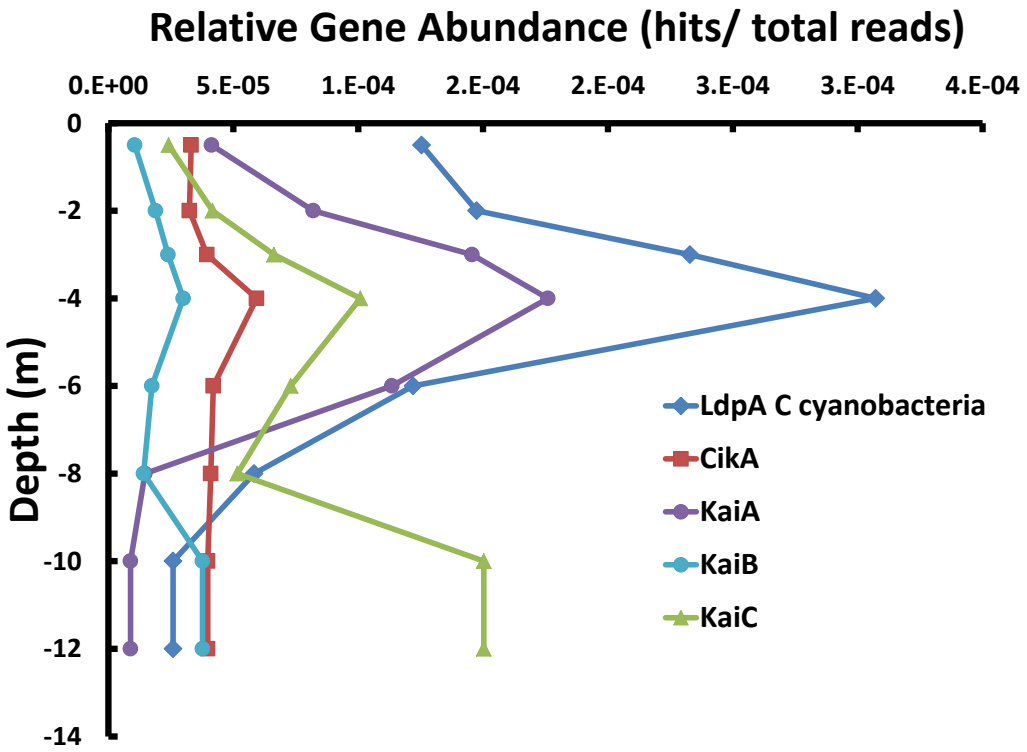

Fig. GB-4. Circadian rhythm genes identified from the metagenomics library from Cast 1.