Prochlorococcus in Context:

Metagenomic Analysis of Sargasso Sea Microbiomes

Reid Pryzant '16 and Claire Ting - Department of Biology, Williams College

Introduction

Prochlorococcus is a numerically dominant photosynthetic prokaryote residing in subtropical and tropical oceans. It plays a key role in global carbon and energy cycles. Although members of the *Prochlorococcus* lineage are closely related, they have evolved significant genetic differences. The question our laboratory is addressing involves how this genetic diversity translates into variation in photosynthetic capacity and the ability to survive environmental stress.

In order to extend our understanding from the laboratory to the open ocean, we have conducted a metagenomic analyses of the Sargasso Sea, where *Prochlorococcus* is often predominant among bacterioplankton. We aim to characterize the Sargasso Sea microbiome and understand how environmental selection shapes *Prochlorococcus* populations at different depths in the water column. We hypothesize that the partitioning of bacterioplankton (and specifically *Prochlorococcus*) populations and functions will exhibit significant differences between near-surface (40 m) and deeper (100 m) waters. These differences should illuminate micro- and macro-scale heterogeneity in key physico-chemical properties and biological interactions in the open ocean. We will attempt to answer these questions through comparative taxonomic and functional analysis along with contrasting cultured and uncultured genomes.

Metagenomics

Metagenomics is the application of genomics to uncultured assemblies of organisms in the natural environment. It is a fairly recent field; the first paper on marine metagenomics was published in 2004 (1).

During August of 2009, the Ting Lab conducted field work in the Sargasso Sea (31° 40.00'N, 64° 10.00'W) on board the R/V Atlantic Explorer. For each environmental sample, we collected over 240 liters of sea water. We subsequently filtered the water (between 0.2 um and 1.6 um) to collect bacterioplankton with Steripak filters (Millipore). All samples were immediately stored in liquid nitrogen and chemical buffers. We then shipped the samples back to our laboratory at Williams College. The Ting lab spent close to year optimizing environmental DNA isolation before extracting the genetic material from our ocean water. We then sent the extracted DNA to the Department of Energy's Joint Genome Institute (Walnut Creek, CA) for Illumina sequencing. In February 2012, the Joint Genome Institute began the process of uploading the metagenomes into the IMG/M annotation pipeline.

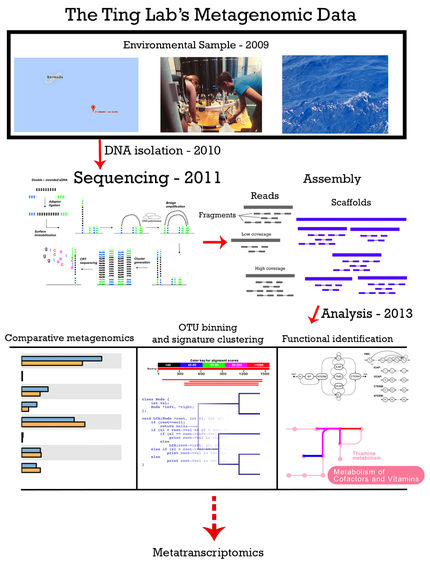

Figure 1: A schematic for the procedure used in metagenomics. It shows key steps from initial field work to the functional identification and taxonomic profiling of over 375,000 proteins. Future work involves metatranscriptomic analysis: the study of gene expression in these samples.

Further analysis was carried out with the MG-RAST pipeline (2), MEGAN (3), STAMP (4), and in-house programs written in Python, Perl, and R. Computation was outsourced to Luis, a server in Jesup Hall running CentOS 6 with 64 AMD 6380 cores at 2500 MHz, 64 gigabytes of memory, and 2.7 terabytes of disk space. Some of the analyses were extremely computationally intensive. Taxonomic profiling of the 100 m dataset, for example, took 10^{13} comparisons and three weeks on Luis.

Feature	40m	100m
Number of scaffolds	45,819	235,606
Total scaffold length (bp)	23,410,463	139,143,246
Average scaffold length (bp)	511	590
Size of .fasta (mb)	24.6	145.4
Number of known protein-coding genes	64,195	331,525
% unknown proteins	19.9	28.9
Average coverage fold	47.2	26.4

Table 1: Summary of the 40 m and 100 m Sargasso Sea data.

Taxonomic Profiling

We hypothesized that the composition of bacterial phytoplankton at 40 m and 100 m will differ significantly. In order to describe the taxonomic distribution of our environmental samples, we identified genes in each scaffold and then assigned them to operational taxonomic units (OTU's) in a process called binning. Gene calling was carried out by comparing each scaffold to the NCBI-nr reference database. Genes were subsequently binned into OTU's with the least common ancestor (LCA) algorithm (Fig. 2).

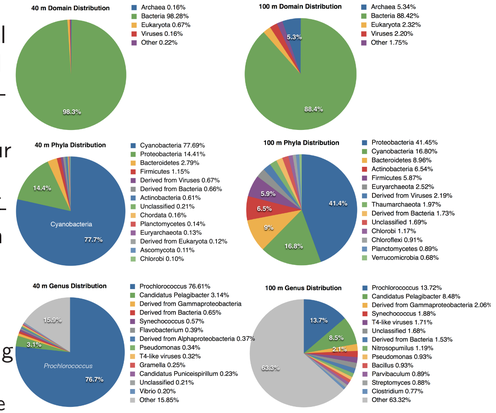

Figure 2: Taxonomic distribution of the 14 most populous groups at the Domain, Phyla, and Genus level. Groups outside the top 14 have been summed and labeled as "Other".

Notably, we discovered that the 40 m microbiome is dominated by Cyanobacteria and, in particular, *Prochlorococcus*. The 100 m microbiome is extremely diverse.

12-strain Comparison

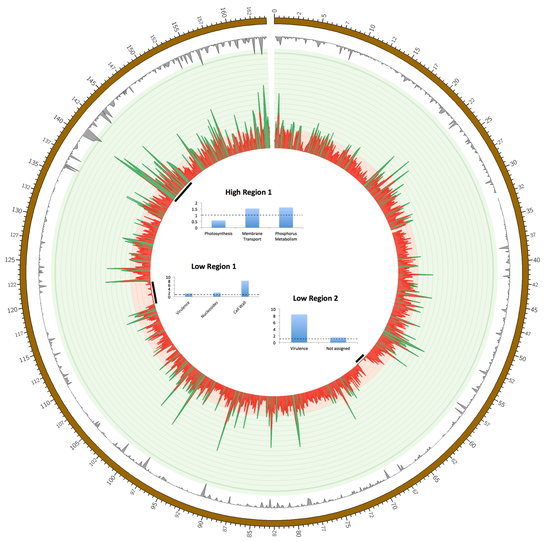

Figure 3: Map of the distribution of shared sequences between MIT 9301 and our environmental samples. This distribution is modeled by inner (40 m) and outer (100 m) lineplots with a 0.5% scale. Position along the genome (bp x10,000) is shown around the edge. Red lines connect points<0.01%. Barplots of odds ratios between expected (genomic) and regional proportions show functional over- and underrepresentation in islands of interest. Hits were filtered by e-score (1e-10), ensuring that highly similar sequences are registered along with identical matches. The number of hits per gene is indicative of abundance and primary sequence conservation (8). "Low Regions" mark the locations of novel, unconserved genomic islands: hotspots of recombination and critical contributers to genetic diversity. "High regions" mark areas of conservation.

In order to illuminate the microdiversity between cultured and wild-type *Prochlorococcus*, the genes of all 12 cultured strains were compared to the environmental scaffolds. The number of subsequences producing significant alignments (hits, with significance determined by e-score) with these genomes were totaled.

Although the *Prochlococococcus* population decreases in abundance with depth, lineages and ecotypes trade dominance throughout the water column. The number of matches between the environmental sequences and cultured strains appropriately reflects this distribution (6) (Table 2). While members of the large clade of recently differentiated lineages have more matches to subsequences of the 40 m sample, members of deeply branched lineages have more matches with the 100 m sample.

Strain	40 m hits	100 m hits
AS9601	23829	18212
CCMP1375	10581	23510
CCMP1986	21301	18434
MIT9211	10292	22007
MIT9215	23532	18201
MIT9301	23832	18374
MIT9303	4977	15879
MIT9312	22781	18488
MIT9313	4953	15401
NATL1A	11569	25672
NATL2A	11627	25762
MIT9515	20918	18229
MIT9301 CORE	17948	16800
MIT9313 CORE	4600	12636

Table 2: The number significant alignments is reported for each strain of *Prochlorococcus*, along with the *Prochlorococcus* core genome.

Functional Characterization

We expect the functional profiles of the water column to reflect micro- and macro-scale physico-chemical and biological characteristics. To describe the functional landscape of our metagenomes, we categorized proteins into heirarchical subsystems. Subsystems are a generalization of the term "pathway." Protein-coding genes were queried against the SEED database for subsystem classification (6) (Fig 4).

- The functional landscapes of the 40 m and 100 m bacterial communities are strongly correlated (r = .962) indicating conservation of broad community functions.
- *Prochlorococcus* at 40 m has a similar functional landscape as its community (d = .89) but *Prochlorococcus* at 100 m is dissimilar to its community (d = 3.07). This is consistent with *Prochlorococcus's* dominance of the 40 m sample.
- In both communities, *Prochlorococcus* accounts for the majority of photosynthetic potential.
- Interestingly, the number of stress response genes associated with *Prochlorococcus* is higher at 100 m than at 40 m.
- Communities loose their photosynthetic potential with depth.

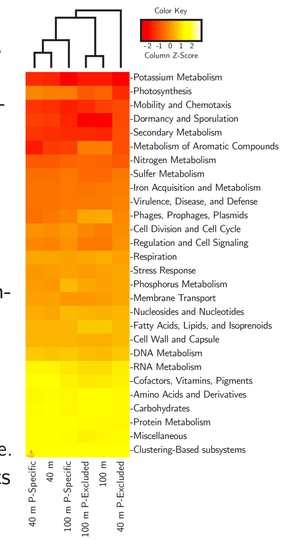

Figure 4: SEED subsystem distribution of all genes, *Prochlorococcus*-specific genes, and *Prochlorococcus*-excluded genes at 40 m and 100 m. Values are standardized log scores and hierarchical clustering is based on euclidian distance (d).

Signature Analysis

To learn more about uncultured, environmental *Prochlorococcus* at 40 m and 100 m, we used sequence signatures to examine inter-genus differentiation (7). We computed tetranucleotide profiles for *Prochlorococcus*-specific scaffolds with a 1-bp sliding window and summed pairs of reverse complementary tetranucleotides. Principal components analysis revealed distinct partitioning of *Prochlorococcus* gene signuatures at 100 m. The smaller group (cluster 1) is closly related to MIT 9303, while the larger group (cluster 2) contains sequences that bear similarity to a broad mix of cultured and uncultured strains.

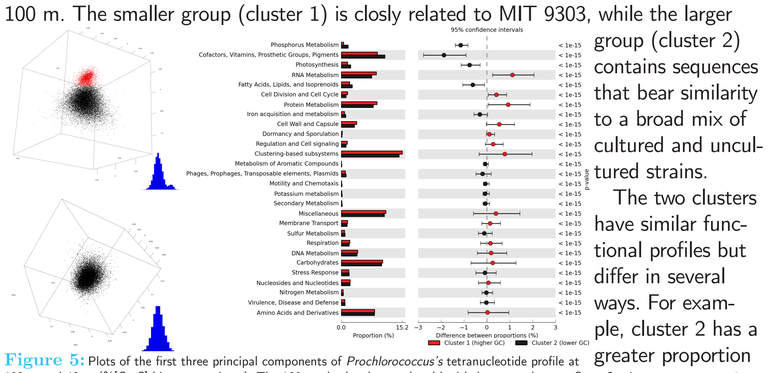

Figure 5: Plots of the first three principal components of *Prochlorococcus's* tetranucleotide profile at 100 m and 40 m (%[G+C] histograms, inset). The 100 m plot has been colored by it's k-means clusters. On the right is a comparative functional analysis between the two 100 m clusters. Wilson confidence intervals and Bootstrap t-tests (100,000 replicates) were computed for each subsystem.

The two clusters have similar functional profiles but differ in several ways. For example, cluster 2 has a greater proportion of photosystem I genes (not shown).

References and Acknowledgements

This work was funded by the National Science Foundation, award number MCB-0850900 to C.S. Ting. I would like to thank professor Ting for her constant help, knowledge and patience, professor Klingenbeg, Adam Wang from OIT, and Jonathon Morgan-Leaman from OIT for access to Luis, and Kathleen for putting up with my antics in lab.

1. Venter, J. C. et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.
2. F. Meyer et. al. (2008) The Metagenomics RAST server – A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9:386.
3. Mitra S, Klar B, Huson DH (2011) Visual and statistical comparison of metagenomes. Bioinformatics btp341.
4. Parks, D.H. and Beiko, R.G. (2010). Identifying biologically relevant differences between metagenomic communities. Bioinformatics, 26, 715-721.
5. Johnson ZI, Zinser ER, et al. (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science 311: 1737–1740.
6. Overbeek R, et al. (2005): The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res, 33(17):5691-5702.
7. Dick G, Andersson A, Baker B, Simmons S, Thomas B, Yelton AP et al. (2009). Community-wide analysis of microbial genome sequence signatures. Genome Biol 10: R85.
ture Biotechnol. 6, 533–538.
8. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. Science 311: 1768–1770.
Photos in Fig. 1: Map: Google Earth. Illumina sequencing: http://www.gsejournal.org/content/figures/1297-9686-44-21-2-l.jpg. COG Markov Model: http://www.quuux.com/stefan/pic/LocApplication.png. BLAST output: NCBI web-BLAST. Assembly: *adapted from* http://genomebiology.com/2009/10/8/R85/figure/F1.