Reconstruction of a Bacterial Genome from DNA Cassettes
Final Report (2005-2011)

Executive Summary

This basic research program comprised two major areas: 1) acquisition and analysis of marine microbial metagenomic data and development of genomic analysis tools for broad, external community use; 2) development of a minimal bacterial genome.

Our Marine Metagenomic Diversity effort generated and analyzed shotgun sequencing data from microbial communities sampled from over 250 sites around the world. About 40% of the 26 Gbp of sequence data has been made publicly available to date with a complete release anticipated in six months. Our results and those mining the deposited data have revealed a vast diversity of genes coding for critical metabolic processes whose phylogenetic and geographic distributions will enable a deeper understanding of carbon and nutrient cycling, microbial ecology, and rapid rate evolutionary processes such as horizontal gene transfer by viruses and plasmids. A global assembly of the generated dataset resulted in a massive set (5Gbp) of genome fragments that provide context to the majority of the generated data that originated from uncultivated organisms.

Our Synthetic Biology team has made significant progress towards the goal of synthesizing a minimal mycoplasma genome that will have all of the machinery for independent life. This project, once completed, will provide fundamentally new knowledge about requirements for microbial life and help to lay a basic research foundation for developing microbiological approaches to bioenergy.

Key Findings

Goal 1: Survey the microbial biodiversity of the world's oceans, construct a freely shared metagenomic database, and leverage it to better understand carbon cycling and to discover genes for biological energy production.

The following specific aims were presented in our 2005 proposal to DOE:

1. Shotgun sequence microbial samples from approximately 150 open-ocean and coastal sites across the Pacific and Indian Oceans. Augment these data with deep sequencing of 16S and 18S rRNA.
2. Develop and refine bioinformatics tools to assemble, annotate, and analyze large-scale metagenomic data, along with the appropriate database infrastructure to enable directed analyses.
3. Undertake specific analyses to better understand carbon cycling and discover genes for biological energy production.

Overview of Accomplishments to Date: *Our sampling of the ocean has created the world's largest size-fractionated anthology of marine microbes including virus concentrates, membrane filters rich with prokaryotes, and eukaryotes less than 20µm in size. This fertile collection of biological material has analyzed as summarized in Tasks 2 and 3 below. DOE funded much of the sample sequencing; sample collection was primarily funded through other resources. In this section, we provide a brief summary of the sampling sites, sequencing data, and our initial sample and data processing activities.*

Our approach comprised five data generation and analysis tasks: (1) Sequencing of samples from the *Sorcerer II* Global Ocean Sampling (GOS) circumnavigation expedition; (2) analysis of metagenomic sequencing results from the GOS circumnavigation, with in-depth focus on Indian Ocean metagenomic sequence and associated metadata; (3) analysis of samples from the targeted coastal collections to study microbial diversity and the influences of microbial community structure on biogeochemical fluxes; (4) develop, improve, and make publicly available metagenomic data analysis tools; and (5) additional sequencing for under-represented and carefully selected samples to enhance the analyses mentioned above.

Task 1: Marine Environmental Sample Sites and Sequence Data

2003-2006 Circumnavigation*:* The *Sorcerer II* Global Ocean Sampling (GOS) circumnavigation expedition collection comprises samples from approximately 150 locations. An analysis of sequence data from the Sargasso Sea pilot study and first leg of the circumnavigation expedition -- a several-thousand km transect from the North Atlantic through the Panama Canal and ending in the South Pacific -- resulted in the 2004 and 2007 landmark publications in Science and PLoS Biology, respectively. The collection represents a wide range of distinct surface marine environments as well as a few non-marine aquatic samples. Most sites have been analyzed for organisms in the 0.1 to 0.8 µm size range with the exception of the Indian Ocean transect; for those sites, we selected five "priority" stations to perform a comprehensive analysis on all organisms from that environment, including larger phytoplankton and viruses.

The Indian Ocean transect of the circumnavigation is a tropical surface water collection of samples, unique relative to the Atlantic and Pacific due to expansive denitrification in the Arabian Sea and associated losses of fixed N causing consistently low N:P ratios. Data analysis for several sites has begun and will continue as described in the Future Work section of this report to investigate relationships among geographical location, nutrient availability, physical forcing (*e.g*. upwelling) and microbial diversity at the phylogenetic and functional level. Key findings are summarized below in Aim 3. The remaining samples from the circumnavigation along with samples from the remaining transects, are the focus of a publication in preparation.

2007 transects: NW Atlantic through Panama Canal to Alaska: This collection is comprised primarily of coastal samples from 55 stations, with an emphasis on near-shore to off-shore gradients and discovery in diverse environments. Some of these novel environments include upwelling waters, hypersaline ponds, high and low pH water, polar ice, mangrove, and tidal areas. A new sample collection technique was implemented which will allow the construction of cDNA libraries from impact filters. Additional metadata were also collected that will further enhance the investigation of biogeochemical pathways.

2007 Collaborative Cruises: Coastal Transects. Samples were collected at 33 stations from near-to-offshore along the California Current Ecosystem and adjacent locales (see Table 1) in collaboration with researchers interested in coastal dynamics. These studies will focus on the unique nutrient gradients associated with coastal regions and the role of microbial communities in biogeochemical flux as discussed in the Future Work section. Analysis on these sites have been concluded and are in preparation for publication or already published.

Antarctica: South Sea Transect: In conjunction with the University of New South Wales and the Australian Antarctic Division, Southern Ocean and Antarctic marine samples were collected aboard the R/V *Aurora Australis.* The Antarctic region is particularly sensitive to changes in climate, and the area sampled, East Antarctica and the Mertz Glacier, is one of the few Antarctic regions where deep thermohaline circulation occurs. In order to better understand the effect of climate change throughout the water column, surface collections were paired with deep water samples, and bottom water was collected on the continental shelf, Adélie depression and abyssal plain. Additional surface samples were collected in nearshore areas, in regions of melting pack ice, near glaciers and around iceberg fields. On the return voyage, a transect of the Southern Ocean was taken from Antarctica to Tasmania, sampling the major physical features of the Southern Ocean, including the Antarctic, Polar, Sub-Antarctic and Sub-Tropical Fronts. A publication on this transect has been submitted and another is in preparation.

East Pacific Rise Deep Sea: Water samples were collected from deep-sea hydrothermal vents and cold deep-sea (~2,500m), outside the influence of hydrothermal fluids. Hydrothermal fluids are rich in reduced chemical compounds that fuel thriving chemosynthetic microbial communities whereas, the surrounding cold deep ocean is a veritable desert where microbes have to efficiently scavenge organic detritus for survival. Size fractionation of microbes and virus concentration was performed *in situ* using a large volume water sampler deployed by the DSV Alvin and the R/V Atlantis.

Tasks 2: Analysis of Sorcerer II circumnavigation results.

We have investigated the fundamental microbial contributions to carbon cycling by organizing our analyses around four interrelated research areas as outlined below: biogeochemical cycling, community structure and function, microbial diversity, and adaptation/evolution. Each focus has been addressed through read-based and assembly based analyses with clear overlap.

Metagenomic Assembly: The high diversity of the microbial communities in seawater presents a barrier to the genome reassembly after shotgun sequencing.

The GBMF funded marine microbe genome sequencing project resulted in nearly 150 genomes for cultivated marine prokaryotes from researchers around the world. One goal of the project was to provide relevant reference genomes for use in metagenomic surveys of the marine environment. These genomes, and all other previously published genomes, were compared to large scale environmental 16S libraries and also used as a scaffold for fragment recruitment (Yooseph, Nealson et al, 2010). Ultimately, the wealth of genomes primarily serves as a example in absence; at 50% nucleotide identity, they only provide a reference for 25% of all metagenomic data (Fig 1) and the environmental abundant 16S sequences were often quite divergent from the closest cultivated representative. Most of the surface marine microbes remain uncultivated. To address this shortfall, we conducted a host of assembly-based studies, using early directed studies to guide a final generate of a genome encyclopedia for the surface ocean.

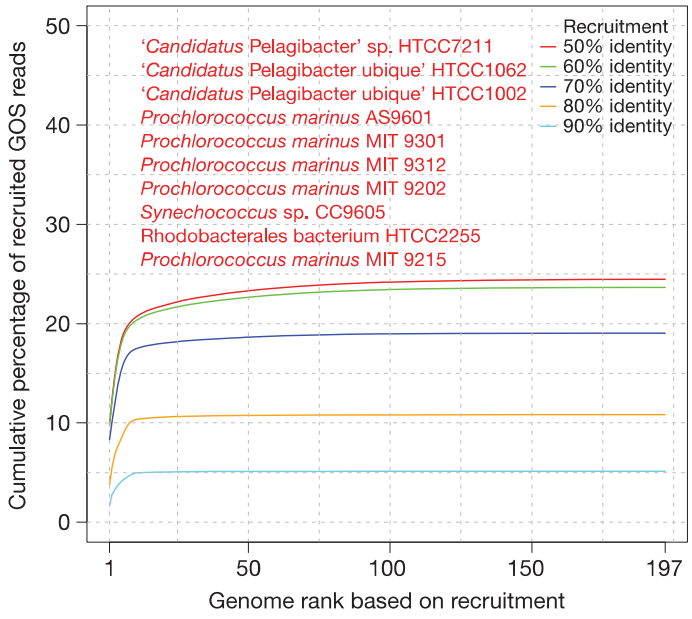


Figure 1. Fragment recruitment of GOS reads at different percent identities to 197 sequenced marine genomes. The x-axis shows genome rank (from highest to lowest) based on the number of recruited GOS reads. The y-axis shows the cumulative (relative to the total number of GOS reads) of those reads that are recruited. The top ten recruiting genomes (at 50% identity) are also listed in the figure and they account for 84% of all recruited reads.

***Directed assembly*: *Prochlorococcus* HNLC strains:** *Prochlorococcus marinus* appears to be one of the few microbial groups where the "great cultivation bias" seems to have been overcome. Numerous strains have been isolated and sequenced, with many of the resulting genomes being among the most abundant in the GOS dataset. However, fragment recruitment analyses revealed a population of *Prochlorococcus* that was dissimilar to both high and low light strains in a series of GOS sites in the South Pacific. Directed assembly of two genomes for this population was aided by already existing *Prochlorococcus* reference genomes. Metabolic reconstructions and an analysis of the global distribution of these two new strains revealed a streamlining in iron utilization and dominance over other *Prochlorococcus* strains in high nitrogen, low chlorophyll regions of the ocean, which have been shown to be iron-limited. The moniker for these regions (HNLC) has been used as the strain identifier for these new genomes.

SAR86: One of the first groups of uncultivated bacteria identified in the Sargasso Sea was the gamma-proteobacterial clade SAR86. Subsequent 16S surveys have verified the ubiquity and abundance (6-10%) of this organism in marine communities. SAR86 remains uncultivated, preventing any interpretation of its role in marine ecology and biogeochemistry. Using early global assemblies, followed by iterative binning and contig orientation, two nearly complete and phylogenetically distinct genomes of SAR86 were assembled. Two separate partial genomes generated using single cell techniques validated the genome structure and content of these metagenomic assemblies. Phylogenomic analyses verified the divegent nature of SAR86; it is a truly unique cluster within the gamma-proteobacteria. Metabolic reconstructions suggest that SAR86 plays an important role in the degradation of lipids and complex carbohydrates, two compounds that the other abundant marine heterotrophic bacterium SAR11 cannot assimilate. As part of both of these projects, methods for genome completeness and size estimation, genome annotation, and metabolic reconstructions were developed. One publication has resulted from this (Dupont et al, online early).

Uncultivated prymnesiophyte: Because microbial communities are highly complex, and many eukaryotes have large genomes and lower gene density than their bacterial and archaeal counterparts, we developed a targeted metagenomics approach to reduce the bioinformatic complexity of the data. Analyses are being done to identify the major gene families for several important groups of eukaryotic microbes. Photosynthetic populations were discriminated based on size and chlorophyll derived autofluorescence characteristics and flow cytometrically sorted from subtropical Gulf Stream waters using a high-speed cell sorter. We applied a series of novel bioinformatics approaches to identify, annotate, and comparatively analyze prymnesiophyte scaffold. Of particular note were two methods for estimating genome completeness and size of partial eukaryotic genomes. Two manuscripts based on this work have been published (Cuvelier et al 2010, Worden, Dupont, Allen, 2011).

Global assembly of the microbial encyclopedia: The entirety of the GOS dataset (~26 Gbp) has been put through a global assembly using the Celera Assembler (allowing for a 14% error rate in read overlaps). The resulting contigs are ~5Gbp in total length and account for ~85% of all GOS reads. Over 1000 scaffolds are over 151 Kbp in length and the N50 is 30,000 bp. By comparison, a recently published marine metagenomic assembly (Iverson et al, 2012) contained 300 Mbp of total assembly with a N50 of 450 bp. These long contigs have been binned using the methods developed for SAR86 (nucleotide frequency, coverage) to create multiple (30+) draft quality genomes. These genome fragments are both abundant and ubiquitious, with most occurring in greater than 30% of the GOS samples. That being said, biogeography of the genome fragments is apparent after sample-contig clustering. A large portion of genome fragments are almost exclusively found in the cold waters encountered in the northeast and northwest US coastlines and the Antarctica transect. Another large set of contigs associated with warm waters. Not surprisingly, several contigs are in nearly every single GOS sample and phylogenetic analyses suggest they are from the SAR11 or *Pelagibactericeae* lineage.
To determine if these draft genomes and genome fragments address our knowledge gap on uncultivated marine microbes, we examined their 16S diversity relative to the high coverage 16S metagenomic clone libraries (Figure 3). Because the genome fragments were assembled and the clone libraries generated full length 16S sequences, the resulting maximum likelihood phylogenies are of high quality. It is clear that the assemblies represent many of the uncultivated OTUs. Preliminary metabolic and abundance analyses have identified several surprising trends. First, a genome for a bacterium in Rhodobacter lineage that appears to conduct anoxygenic photosynthesis and carbon monoxide production was assembled; relative to all other Rhodobacter strains this one is ~10X more abundant. A similar story was observed for methylophillic beta-proteobacteria, and we expect that a comparison to the genome of the less abundant but cultivated relative will be informative. Additionally, several contigs of clear *Prochlorococcus* origin were only found in the cold (<8^{o}C) waters of the Antarctica transect. A publication on this dataset is in preparation and it is expected that this genome encyclopedia for the surface ocean will aid in cultivation efforts, the interpretation of metagenomic data, and the study of gene and operon biogeography.

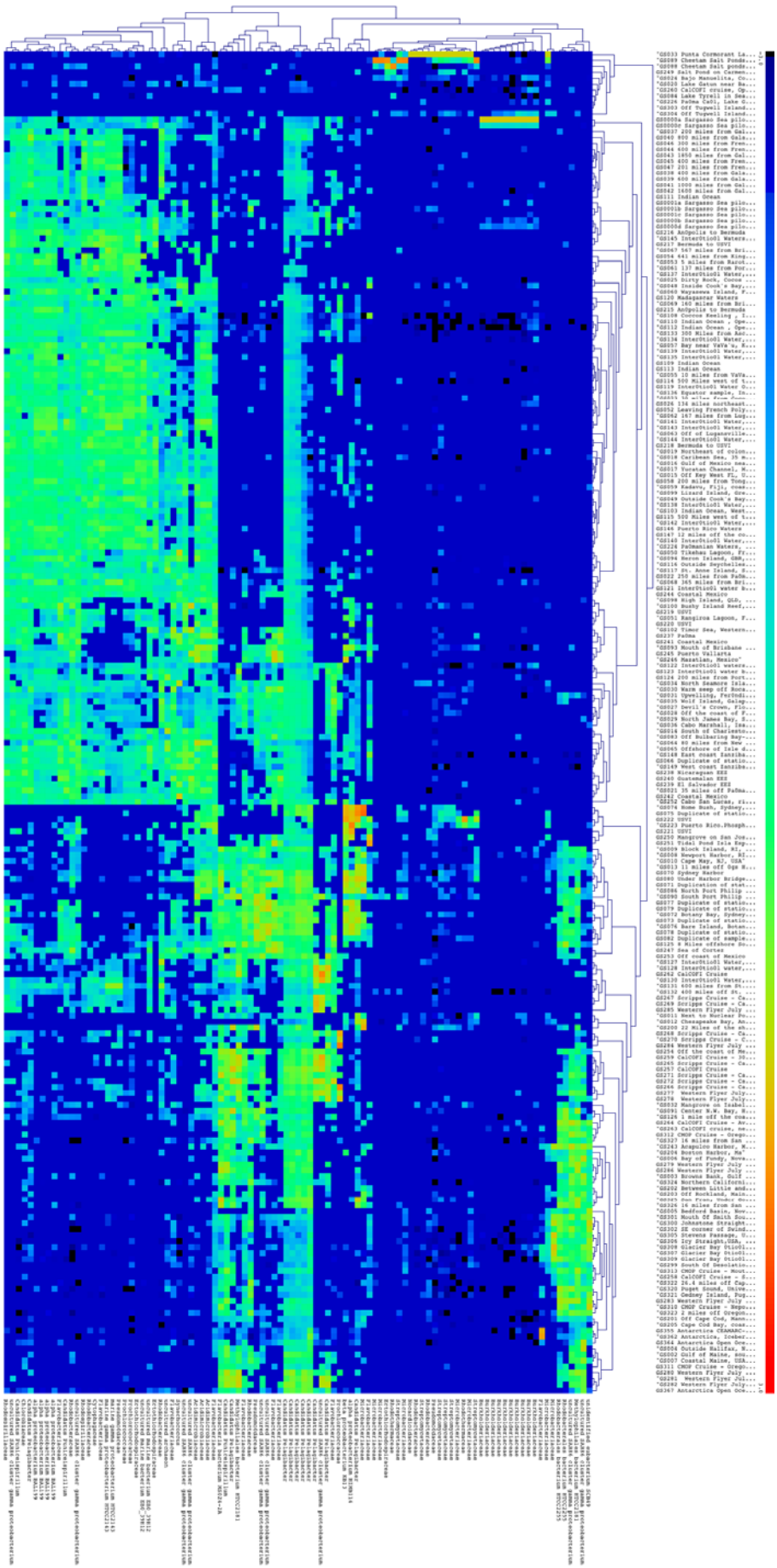
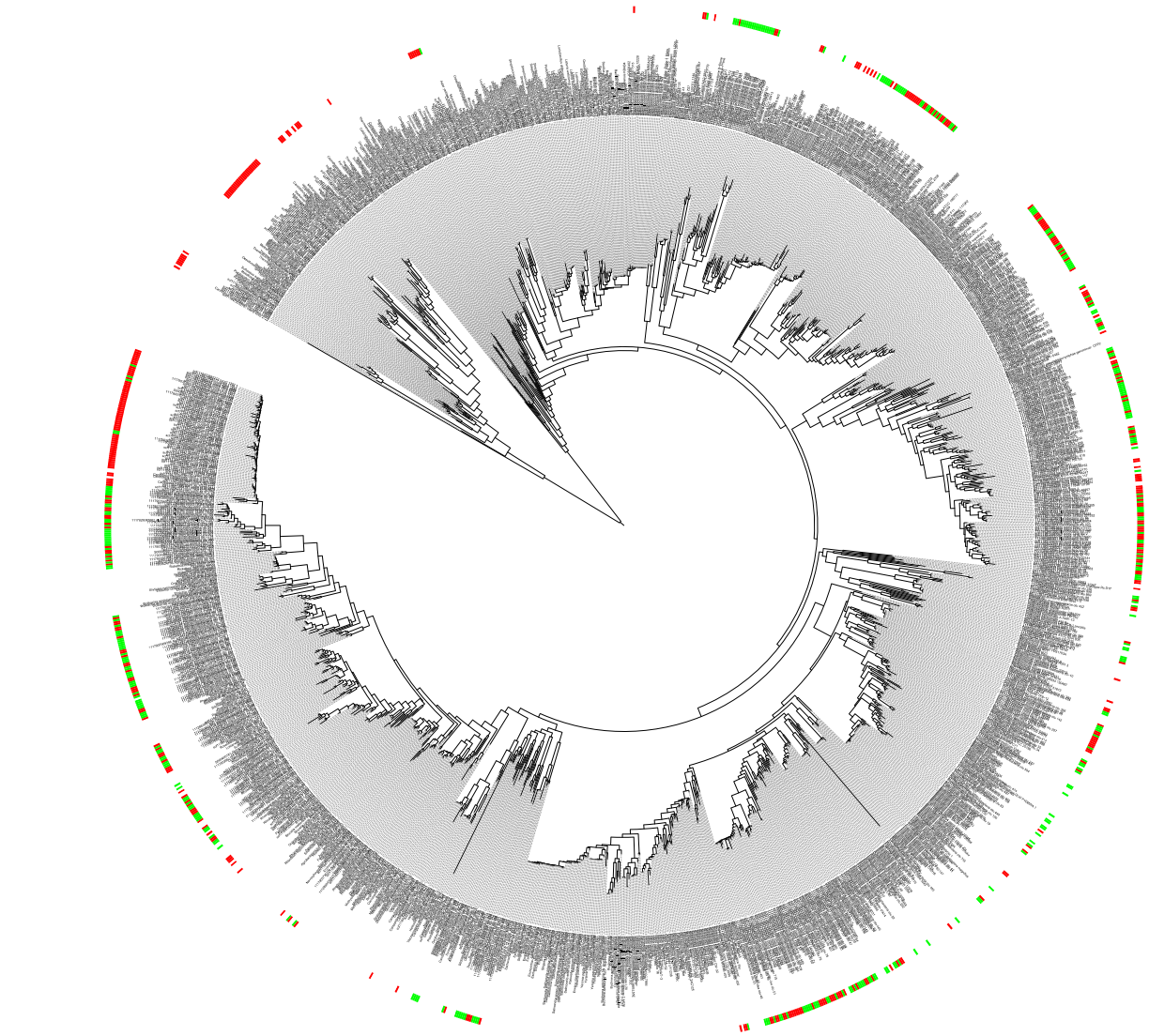


Figure 2: Metagenomic recruitment of the 100 largest assemblies from all GOS sites. Each column represents an assembly >500 Kbp in length, while each row is a GOS site. Each assembly was used to recruit the reads from the metagenome present at each site. Within each assembly (i.e., column), blue indicates relatively more reads were recruited from a particular site (row); green fewer; and yellow to red, the fewest. These values were used to cluster assemblies with similar distributions and sites with similar composition. The sites group relatively well with temperature, as the majority on the top are warm water sites while those on the bottom are cold water sites. The assemblies that recruited reads from all sites are *Pelagibactericeae.*

Figure 3: 16S tree of genomes, marine 16S clone libraries, and metagenomic assemblies. In the figure below, uncultivated OTUs are represented by a green hash while 16S sequencing generated from the global GOS assembly are noted in red.



Read based analyses

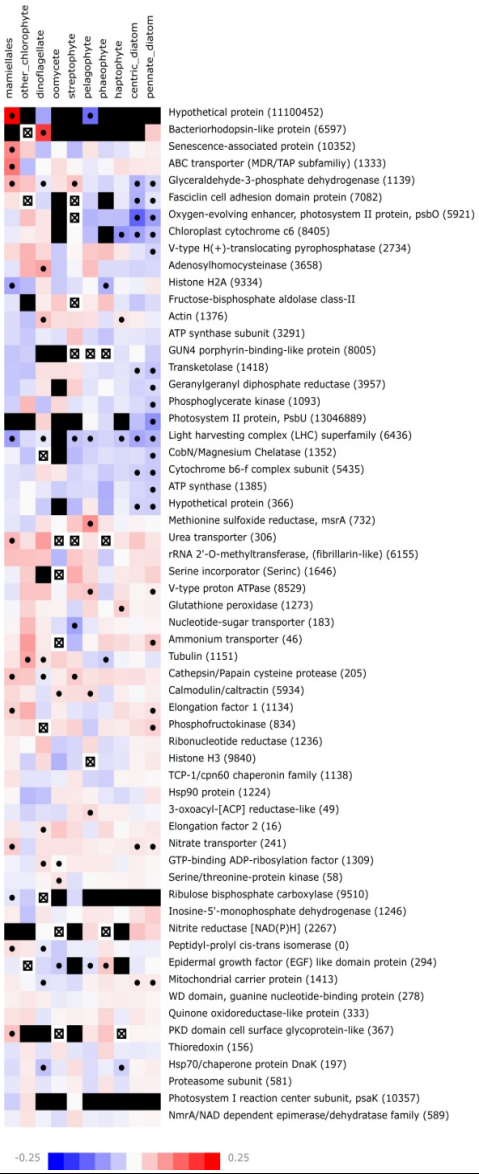
Keystone enzymes. In order to link microbial phylogeny with putative gene function and to specifically associate the flux and transformation of key elements (C,N,S, Fe and other trace metals) with particular organisms, we have selected a suite of around 50 genes that encode "biogeochemical keystone" enzymes. These genes encode for proteins that control important carbon assimilation, degradation and transformation reactions that are important for regulating major element cycles. In addition, we are also interested in exploring the diversity and distribution of a variety of stress response genes. To characterize the keystone enzymes within samples we: 1) Interrogate protein clusters representing previously published GOS data (Rusch, Halpern et al. 2007) (referred to here as Phase I) and GOS Indian Ocean data to identify and quantify the number of keystone enzymes present within samples and normalize them according to estimates of the total number of genomes present; 2) Analyze the distribution and diversity of keystone enzymes by constructing phylogenetic trees of keystone genes that are

present in microbial reference genomes and mapping metagenomic data to the trees; 3) Explore operon context, structure and diversity by cataloging the neighboring genes in reference genomes operons as well as metagenomic mate-pairs; and 4) Evaluate the contribution of viruses to the diversity and distribution of keystone enzymes (See details on progress under the Adaptation & Evolution section below).

In order to generate computational infrastructure necessary to develop pipelines appropriate for these tasks we have done the following:

Cluster based annotation

We used the protein clustering resource developed for GOS Phase I data analysis (Yooseph, Sutton et al. 2007) and which is periodically updated with new data (Yooseph, Li et al. 2008). Functional annotation of protein clusters was performed and utilized to increase annotation coverage of GOS peptides and to describe the distribution of functions across samples. Protein clusters containing genes from both public sequenced genomes and GOS metagenomes were annotated based on the presence of similar annotations by multiple peptides within the same cluster. Functional annotations of clusters are based on: NCBI gene, COG, EC, GO, Pfam, TIGRfam, TIGR role, KEGG ortholog, module, pathway. A general statistical test is used to determine functional assignment of any of these annotation types to a cluster based on a Fisher p-value. This p-value is modeled after the hypergeometric distribution and describes the likelihood of a particular annotation being associated with a cluster at random given the size of the cluster and the scope of the given annotation. A tool to explore the content of protein clusters and all the associated functional annotations with p-values was developed. This tool can also be used to visualize the distribution of clusters across GOS sites, and across taxonomic groups.



Correlations between variations in functions at different levels and sample metadata are being made that leverage the protein clustering for higher coverage of annotations. An example analysis resulting from quantification and annotation of protein clusters in metatranscriptomic data is given in **Figure 4**.

Figure. 4 Heat map representation of gene transcription in Monterey Bay phytoplankton communities in response to nitrogen level based on GOS protein clusters. Cluster ID(s) are given in parentheses and annotations are derived from cluster-level annotations derived in this project . Represented clusters had significant expression levels ($P<0.01$) for reads that mapped to at least two different reference taxa. All clusters indicated were represented by differential expression ($P<0.05$) for at least one taxonomic group. The red and blue coloring represent relative differential expression levels in replete and deplete nitrogen conditions respectively. A box with an X or black coloring indicates that transcripts for a given cluster were not detected or are absent in certain target taxa respectively.

Proper normalization of metagenomic samples is required so that different functional and taxonomic profiles are comparable. The number of sequence reads from a sample is not an adequate proxy for the number of genomes sequenced in a sample because it does not take into account variations in genome size or gene copy number.

Sequenced genome equivalents have traditionally been estimated by a single gene (recA), which is highly conserved and usually in single-copy. Multiple sets of conserved core bacterial genes have more recently been used (Wu and Eisen 2008). We use a set of HMMs based on 31 conserved, single-copy, core bacterial genes to estimate bacterial genome equivalents to normalize metagenomic samples prior to analysis. Estimation of genome equivalents is based on a model fit to simulated genomic and metagenomic samples of complete bacterial genomes. These same 31 core HMMs were also used to construct a phylogenetic tree of life of more than 800 bacterial reference genomes, which is used as a reference to efficiently assign taxonomies to new metagenomic reads. Metagenomic samples are scanned for predicted bacterial peptides that match one of the 31 core marker genes, and taxonomic assignments are made for these peptides yielding a normalized taxonomic profile that is comparable across samples. For each core peptide a subset of potential trees are compared by maximum likelihood methods. The subset of potential trees searched are the best placements of the query peptide to individual nodes on the reference tree at various depths by comparison to precomputed position-weight-matrices built from reference sequences below each node. Tools have been built to automatically search metagenomic sequences for core bacterial marker genes, estimate genome equivalents for normalization, and perform taxonomic profiling based on these core genes. A tool has also been built to visualize the normalized distribution of metagenomic samples across the tree of life and produce publication quality annotated large circular trees.

Estimating genome size from sequenced genomes

A simple linear relationship exists between the number of genes in a bacterial genome and the size of that genome in nucleotides. We developed a tool to calculate the average genome size within a particular filter sample. This new tool is being used to normalize genomes across samples and sites. A linear regression was performed over all complete bacterial genomes in NCBI and the resulting model used to estimate average bacterial genome size from GOS samples (**Figure 5**). A subset of predicted peptides was extracted for each sample that was assigned to the domain Bacteria using the Automated Phylogenetic Inference System (APIS), an in-house tool for phylogenetic and taxonomic classification and analysis of genomes. Genome equivalents were estimated on this subset of peptides based on core marker HMMs. The average number of peptides per genome was fit to the linear model to estimate average bacterial genome size of each GOS sample. Results show the correlation between filter and genome size; the larger the filter the larger the genomes found in a sample.

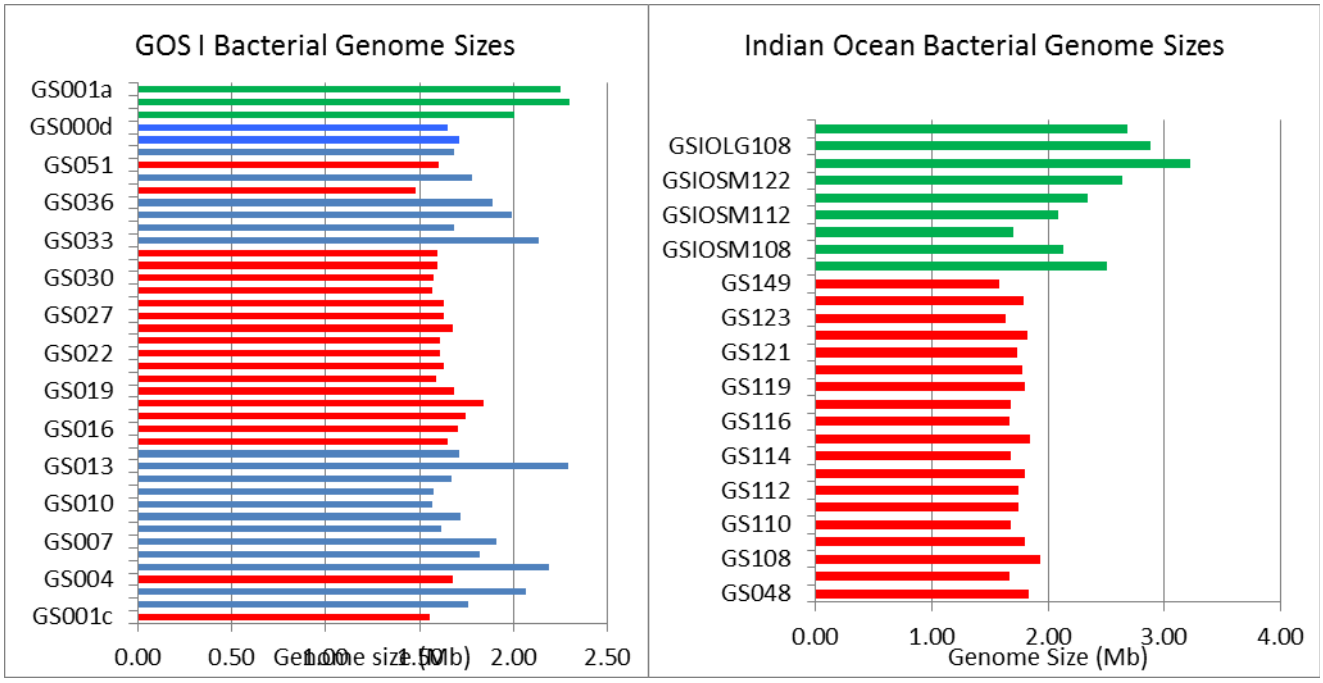


Figure 5. Estimated Average Genome Size for GOS Sites/Filters Phase I and Indian Ocean
Green = 0.8 and 3.0 filters; **Blue** = 0.1 filters meso- and eutrophic sites; **Red** = 0.1 filters oligotrophic sites

Community Structure and Function

Considering the vast nature of our sample collection, the community structure and function of the microbial communities represented by these samples will undoubtedly vary considerably in response to both geography and resource availability. We anticipate that the majority of marine ecosystems share a common metabolic signature. However, the microbial phylotypes that encode this signature are likely ecosystem-specific. We hypothesize that the rare organisms (i.e. the long tail of diversity) code for unique metabolic capabilities that are essential to the overall function and balance of a particular marine ecosystem. To test these hypotheses we are working to: 1) Examine phylotype structure and variation among sample sites and investigate habitat-specific gene acquisitions. Systematic analyses of the consistency of the gene repertoire of the same phylotypes across different habitats will provide information on the impact of habitat-specific gene recruitments and acquisitions; and 2) Identify the rare organisms in our sample collections. Total phylotype inventories are computed by total phylogenomic profiling (corrected for horizontal gene transfer). The biogeography of phylotypes and associated genes that make up minor components of communities are being analyzed in more detail.

In order to accomplish these tasks we have implemented core HMM and APIS based diversity estimators to substitute for ribosomal RNA sequences, increasing the number of phylogenetic markers and the depth of phylogenomic profiling. Diversity metrics across GOS samples and size classes better describe the range and spread of taxonomy **(Figure 6)**, particularly in relation to environmental variables or niche.

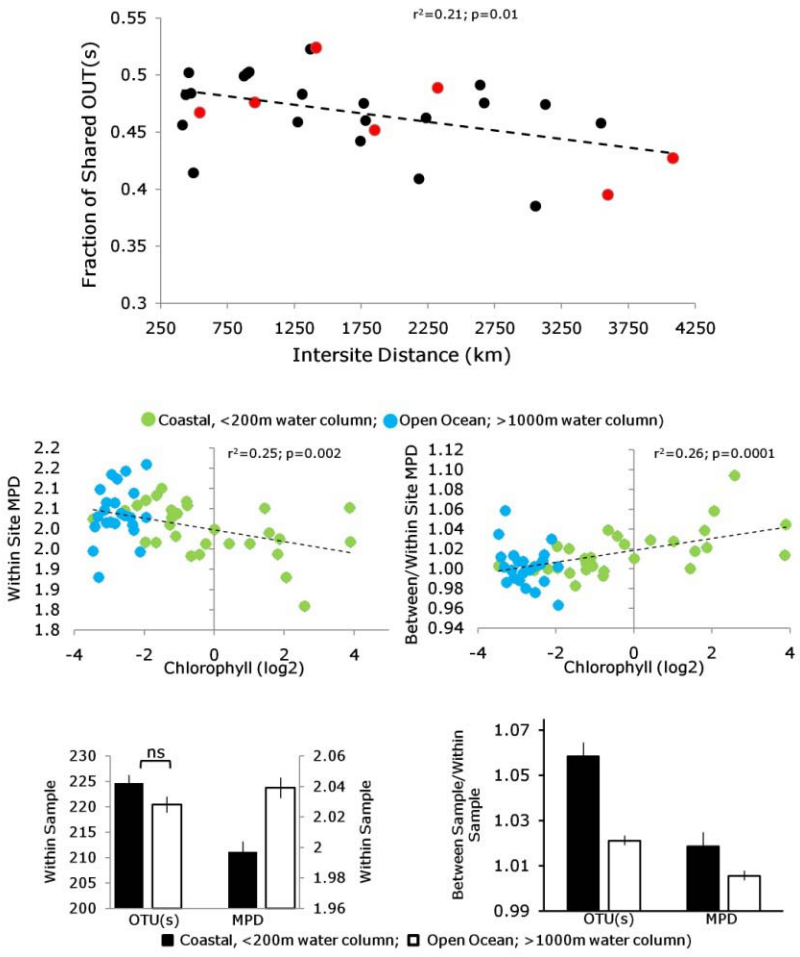


Figure 6. (**Top**) Patterns of OTU overlap based on phyolgenetic placement of bacterial marker peptides onto tree of life across Indian Ocean transect. The negative relationship between geographic distance (x-axis) and degree of shared OTU(s) in pair wise comparisons between samples suggests spatial and biogeographic segregation among microbial populations. (**Middle**) Mean pairwise genetic distance (MPD) is negatively correlated with primary productivity and generally higher in open ocean compared to coastal samples. Also the level of between to within sample MPD diversity is higher for coastal compared to open ocean samples suggesting that sample to sample

Based on findings related to significant differences in community diversity profiles between coastal and open ocean sites we further compared these sites based on transporter abundance and function and carbon and nitrogen content of the predicted proteome (**Figure 7**).

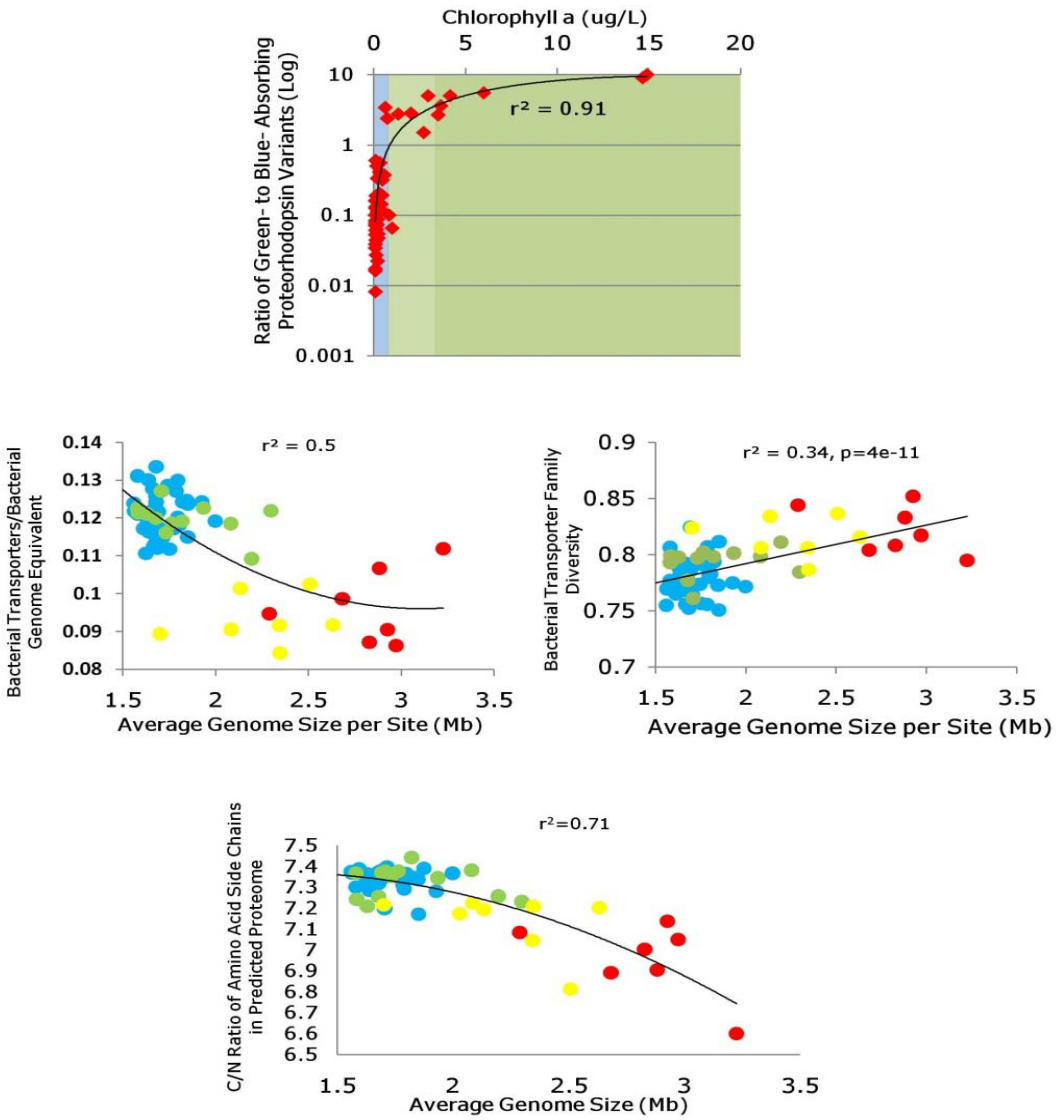


Figure 7. (**Top**) Genome size normalized abundance of green to blue light absorbing variants of proteorhodopsin (**Middle Left**) Bacterial genome size normalized abundance of bacterial transporter proteins. In relation to genome size. (**Middle Right**) The diversity (Simpsons) of bacterial transporter families as a function of genome size. (**Bottom**) Relationship between carbon/nitrogen content of predicted metagenomic bacterial proteomes in relation to genome size. Blue circles represent 0.1 um open ocean filters, green represent coastal 0.1 um filters, and yellow and red indicate samples collected on 0.8 and 3.0 um filters respectively.

Viral Communities

Viruses (primarily bacteriophages) are the most abundant biological entities on our planet and greatly influence the population biology of their microbial hosts. Although viruses exist in every environment, their influence on the global geochemistry is the most profound in the oceans. It is estimated that viruses kill 20% of the microbial biomass in the oceans per day, as reviewed in (Suttle 2007). To more fully understand the implications of virus-host interactions, we are trying to identify the putative microbial hosts of viruses in marine ecosystems. In the absence of direct co-cultivation efforts, this can be extremely challenging. However, we are in an ideal position to address this challenge due to our extensive metagenomic data collection.

To infer putative host ranges for viruses in the GOS samples we developed the first classifier that predicts microbial hosts for the bacteriophages based on the metagenomic sequence fragments. The new version trains Interpolated Context Models (ICMs) on all bacterial genomes in NCBI RefSeq, and assigns viral metagenomic contigs (at least 5Kbp long) to the putative hosts with the highest scoring models. We also compiled a benchmarking database of known phage-host pairs from GenBank records consisting of fully sequenced genomes. On a non-redundant set of 60 phage-host pairs (one pair per microbial genus) with 5Kbp fragments, the ICM based classifier demonstrated a 50% accuracy when selecting hosts among all NCBI RefSeq genomes (n=1530), reported at (Tovtchigretchko A. Composition based and CRISPR based prediction of bacteriophage-host association. 2010, 3rd CRISPR Research Conference, Berkeley, CA). The classifier is available in source code and as a public Web server (http://mgtaxa.jcvi.org/). We applied the classifier to GOS viral contigs (Williamson et al, manuscript in preparation). We also used 3D Principal Component Analysis (PCA) plots of sequence composition features for the visual analysis of specific virus-microbe samples in GOS data and observed a clear association between viruses and putative microbial hosts in several cases (**Figure 8**).

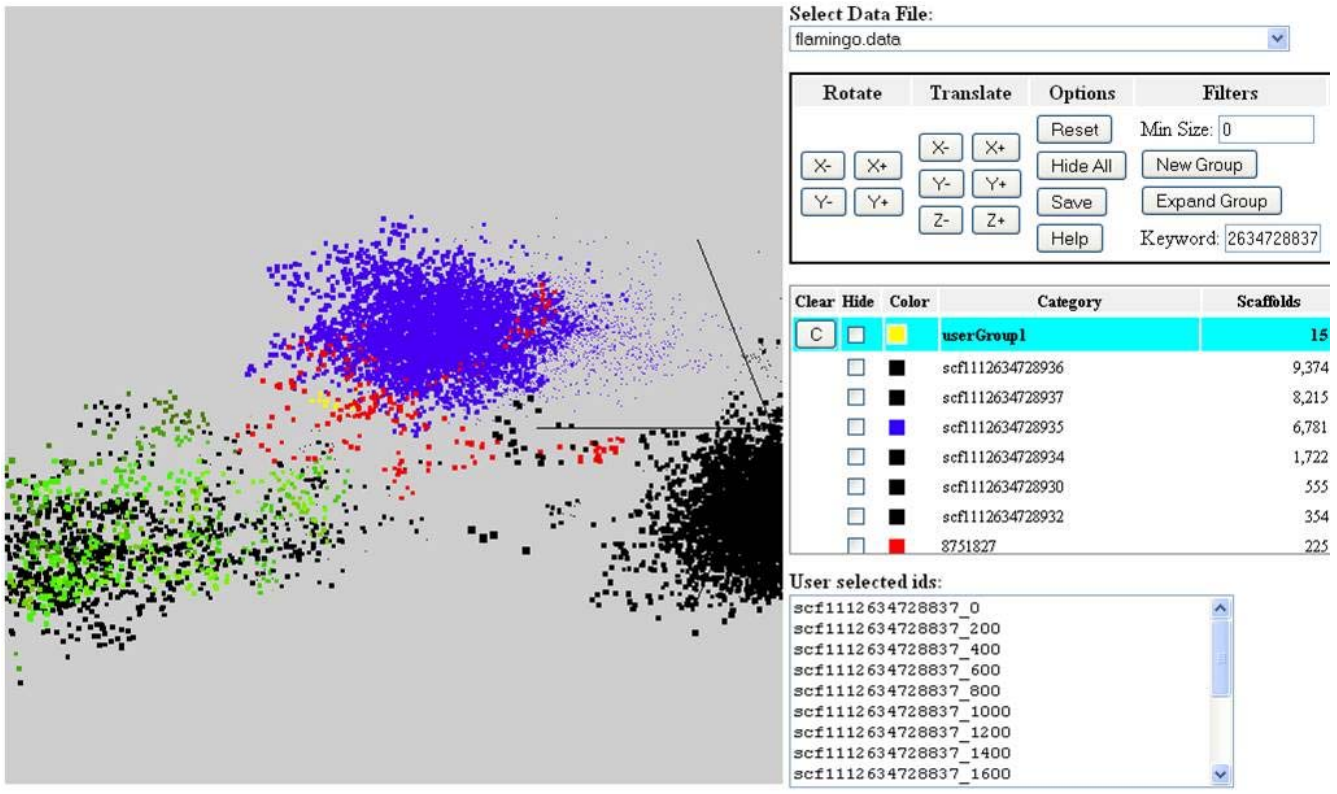


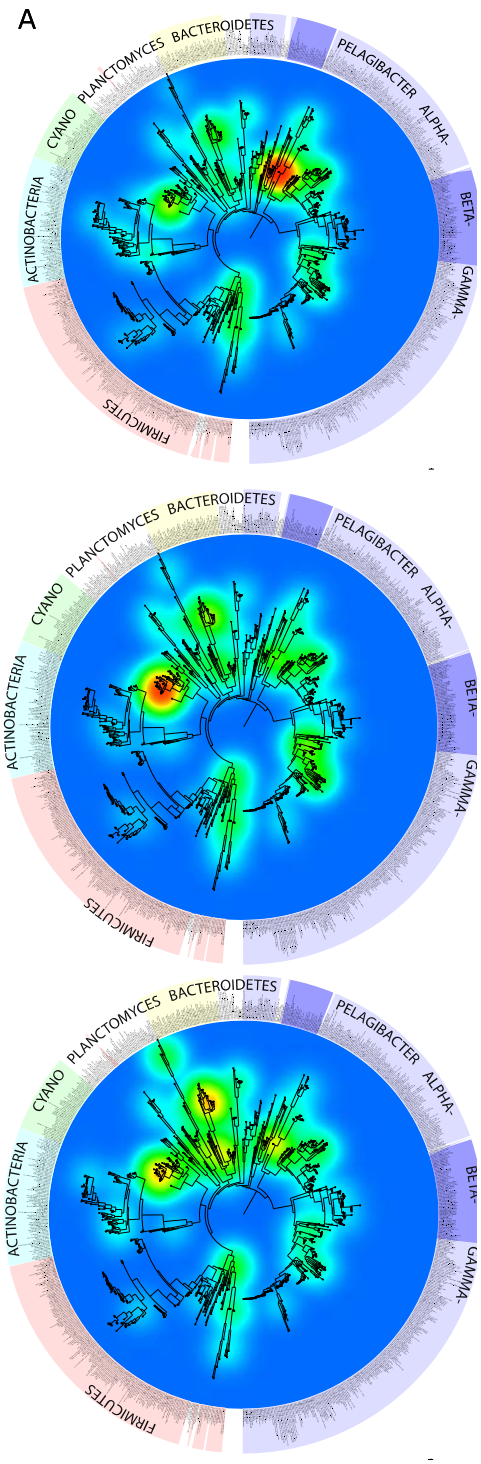
Figure 8. 4-mer PCA of siphophage (red) and bacterial scaffolds (other colors) from the Flamengo Pond metagenome. Blue color represents one of the most abundant bacterial genomes found in this dataset.

Microbial Diversity:

Bacterial Diversity

Methods have been developed to measure the average distance across the tree of life between all pairs of predicted bacterial peptides as a measure of taxonomic diversity or richness. We completed an analysis of diversity across Indian Ocean GOS samples and filter sizes to better describe the range and spread of taxonomy (**Figure 9**). This measure differs from counting Operational Taxonomic Units (OTUs) because it makes use of diversity on all taxonomic levels rather than only below a certain cutoff and would be sensitive to comparisons of multiple OTUs within the same phylum vs. spread out amongst multiple phyla, for example.

Figure 9 Taxonomic diversity of GOS Indian Ocean data by filter sizes life at 5 priority stations. Microbial taxa based on core HMMs mapped to Tree of Life. Top (0.1), middle (0.8), bottom (3.0)



Eukaryotic Diversity

The GOS data presents several unprecedented opportunities for the analysis of genetic diversity of marine microbial populations. Our most recent analyses have involved looking at the phylogenetic composition and diversity of eukaryotic-specific protein families in order to provide information on genomic elements enriched in genomes of eukaryotic plankton relative to bacteria. A limited Sanger sequencing survey of eukaryotic size-class filters (0.8 µm and 3.0 µm) collected in the Sargasso Sea and during a transect of Sorcerer II through the Indian Ocean was conducted. Approximately 50,000 Sanger reads (~35 Mb) were obtained for nine different filters. Preliminary analyses of these data reveal a variety of protein families that are significantly enriched on or specific to larger pore-size filters. The majority of these protein families display a strong eukaryotic phylogenetic signal.

Other large classes of eukaryotic-specific protein families putatively encoded by genes uncovered through this analysis include mysosins, fatty acid synthases, small GTP-binding proteins, seven transmembrane helix receptor proteins, proteases, G protein subunits, multi-drug resistance proteins, and others. An in-house pipeline (APIS) for phylogenomic profiling suggests that certain protein families are specific to particular lineages of microalgae. We have developed methods for gene calling, annotation, and assignment of genes to taxa.

Two of the most abundant gene families on the larger size-class filters belong to stramenopile-like gag-pol polyprotein and reverse transcriptase domains. Further analyses indicate that these proteins belong to copia-type LTR retrotransposons (LTR-RTs). Intact LTR-RTs are large open reading frames (ORFs) (around 6 kb) and preliminary calculations, based on our Indian Ocean data, suggest that copia-type LTR-RTs might comprise one of the most abundant ORFs in the oceanic surface environment. Whole genome analysis of the *T. pseudonana* and *P. tricornutum* genomes had earlier established that LTR-RTs are the most abundant transposable elements (TEs) inhabiting these genomes (Bowler *et al.,* 2008). LTR-RTs are thought to be extremely important contributors to eukaryotic genome evolution by inserting into genes or genetic regulatory elements, thereby disrupting gene function, altering levels of gene expression, triggering chromosomal rearrangements, and adding or subtracting from the physical size of a host genome (Kumar and Bennetzen, 1999). LTR-RTs can be important drivers of population diversification, as they are often responsible for forging haplotype diversification and ultimately limiting sexual recombination. LTR-RTs are also genetic mediators of physiological acclimation to stress conditions (Biémont and Vieira, 2006).

Also, we have initiated several metatranscriptomics projects in order to identify genes highly expressed in various marine phytoplankton taxa. A series of experiments following bloom development and decline in mesocosms incubated with Monterrey Bay upwelling water and sampling transects across the Southern California Bight and in Puget Sound have been completed, metatranscriptomic cDNA samples have been prepared and sequenced. Data analyses are currently underway.

Functional Diversity

We have adapted and tested informatics pipelines for assembly scaffold analysis to look at genome regions particular to specific environments, and are applying these pipelines to examine intra-phylotype diversity as compared to inter-phylotype diversity to determine if certain gene families and operon structures are restricted to particular phylotypes. By doing so, we introduce the concept of "*operon biogeography*" whereby operon structures from complete microbial genomes are used to retrieve and analyze operon-level information from assembled marine metagenomic data. We hypothesize that variation in operon structure across relative resource gradients in marine ecosystems will reveal significant shifts in functional diversity that are not evident via analysis of traditional community structure indicators (*e.g*. 16S rRNA and other molecular and taxonomic markers). To characterize operon biogeography, we are working to:

1. Determine if certain gene families and operon structures are restricted to particular phylotypes by evaluating scaffold ORF composition. Protein clusters are being examined for co-localization on phylotype specific scaffolds and contigs and their relative distributions in the GOS data set will be examined to collect information of core and pan operon components.

2. Perform functional annotation of phylotype-specific operons via Gene Ontology for abundant microbial taxa.

3. Perform fragment recruitment to reference genomes to identify hotspots of hypervariability (insertions/deletions).

Viral Contribution to Microbial Diversity

Viruses behave as global reservoirs of genes. Viruses, particularly bacteriophages (phages) facilitate horizontal gene transfer between prokaryotic host cells. Phages also notably adopt metabolic genes from their hosts. Expression of these co-opted genes during the viral replication process is hypothesized to extend the lifespan of the host; ultimately increasing replication efficiency and the overall fitness of the virus. These genes enter the global pool of genetic information and become available for transfer back to host organisms and/or to co-infecting viruses; thereby influencing the diversity of microbial populations. Prophages (i.e. integrated temperate phages) can also account for a large proportion of the sequence variation between closely related strains of bacteria and subsequently may contribute to the genomic heterogeneity of marine microbes. To quantify viral contribution to microbial diversity, we are:

1. *Determining how viral genes (originating from the viral fraction of samples and those identified as viral in the larger size fractions) influence the gene catalog of microbes within GOS samples.* (See information under Adaptation & Evolution below)

2. *Identifying putative prophage and prophage fragments on all assemblies larger >10 KB. Since each scaffold is given a taxonomic assignment, we will be able to evaluate how prophages contribute to the diversity of various microbial populations.* Protein clusters containing phage-related tyrosine recombinases (i.e. integrases) were identified and the phage integrase sequences were mapped to co-assemblies of microbial and viral data (scaffolds). A total of 1,205 scaffolds containing phage-like integrases were identified. In order to examine the genomic context of integrase sequences and to determine if viral-encoded cellular genes were encoded by prophage, a contig/scaffold graphical display tool was developed, based on the current version of GBrowse (a publicly available genome browser). This tool graphically provides information on the position of the integrase on the scaffold, neighboring genes and their associated annotations and assigned taxonomy and the overall taxonomic assignment of the scaffold. Scaffolds demonstrating an enrichment of known phage-related genes coupled with cellular genes are currently under more extensive analysis.

Adaptation and Evolution

The survival of microbes in an environment is ultimately dependent on their ability to adapt to local conditions. The plasticity of microbial genomes is a function of mutation, gene acquisition through horizontal gene transfer and gene loss; a collective set of events that ultimately guide the evolution of microbial lineages. Microbes will inevitably adapt to the challenges presented by a particular niche in the marine environment, yet we do not understand to what extent the microbial genomic repertoire reflects local adaptation. We hypothesize that variable parts of the genome (i.e. the accessory genome) determines local adaptation and that these regions may contain genes that increase an organism's fitness in those environments. To test this hypothesis we have developed informatics techniques to find regions on assembly scaffolds that are sample or environment specific.

We examined the gene repertoires from several GOS sites using a combination phylogenomic profiling and associated recruitment from taxonomically identified contigs and scaffolds. Our preliminary results show only a small portion (1% or so) of the assembled sequence seems to be region specific. The results make sense given what we know about the sampling sites and the datasets. This success of this approach is linked to the amount of available sequence and the abundance of the individual organisms. Therefore, preliminary investigations have been focused on comparing the gene content of the most abundant and widely distributed marine microbe, *Pelagibacter,* across warm and cold water gradients. Several candidate regions that are differentially present across this gradient have been identified. We are in the process of verifying these discoveries and are analyzing their gene content to determine their significance. In the future we plan to extend this analysis to a wider range of organisms and to explore other environmentally and ecologically important gradients.

We are continuing to analyze the accessory genome using the techniques described above to: (1) Evaluate the pan-genome (the combined and largely conserved gene set for major marine microbial ribotypes) expansion for various microbial taxa and determine its contribution to regional adaptation; and (2) Characterize the accessory genes of the dominant microbial taxa identified in our samples and determine the source of the accessory genes. Large scale comparative genomic analyses within and between reference databases will provide information on core and accessory genome components for different groups. The generation of a genome encyclopedia by metagenomic assembly has allowed us to extend these analyses to clades of uncultivated bacteria.

Viruses Role in Microbial Evolution

Viruses play a significant role in the evolution of their microbial hosts due to their ability to efficiently shuttle genes from one cell to another. As previously mentioned, we are becoming increasingly aware that viruses acquire metabolic genes from their hosts as part of their replication and survival strategy. These genes may also play a larger role in the adaptation of microbes to local conditions (e.g. nitrogen utilization genes in nutrient deplete environments). By and large the majority of viruses that participate in this process appear to be lytic rather than temperate phages (Sullivan et al. 2006); very little is known about the occurrence of cellular metabolic genes within marine prophages. In order to explore the influence of viruses on the adaptation of marine microbes we are:

1. *Metabolic profiling of viral genes (from viral metagenomes and larger size classes) to characterize them in the context of cellular metabolic pathways.* Protein cluster annotation was used to identify those clusters that 1) mapped to metabolic pathways and 2) contained viral sequences from the viral metagenomes and the larger size classes. Viral sequences mapped to both expected and unexpected pathways. The expected pathways included amino acid, carbohydrate, nucleotide and glycan metabolism as well as replication/repair and transcription/translation. The unexpected pathways included cell growth and death, production of cofactors and vitamins, energy production, folding, sorting and degradation, lipid metabolism, membrane transport, secondary metabolite metabolism and isoprenoid biosynthesis.

2. *Identification of the types of viruses (e.g. lytic or temperate phages, and eukaryotic viruses) acquiring metabolic genes and potential genomic sources via phylogenetic analyses and mapping to prophage regions of assemblies*. Phylogentic analysis of viral sequences that appear to be linked to interesting metabolic pathways was used to determine the diversity and taxonomic distribution of viruses that may participate in these cellular functions. Viral cellular genes were mapped to scaffolds in a similar fashion to the collection of phage integrase sequences (described above) in order to determine the genomic context of the viral genes and the modified GBrowse tool used to graphically display the contextual information. The taxonomic assignments (assigned by APIS) of the viral genes on the scaffolds was used in order to determine the types of viruses that carry cellular genes. Evidence of co-localization of phage integrases and viral cellular genes also indicates temperate phages carry cellular genes in addition to lytic phages.

3. *Identify phage integration via CRISPR Analysis.* Viruses also instigate rapid evolution of microbial hosts in the form of CRISPR systems. CRISPRs, or Clustered Regularly Interspaced Short Palindromic Repeats are widely distributed in prokaryotic genomes and are believed to constitute an anti-viral immune system. CRISPR arrays therefore represent a kind of phage predation-driven molecular clock, likely the most unambiguous clock derivable from Global Ocean Sampling metagenomic data. We implemented a pipeline to analyze CRISPRs and associated genes (CAS genes) in metagenomes and complete genomes. Our pipeline aims at establishing a connection between the viral and microbial metagenomic fractions by finding sequences in the viral metagenome that match CRISPR spacers from microbes. It was applied to both metagenomic GOS datasets (viral and microbial fractions) reported at a conference (Williamson 2009), and Moore genomes (included in the paper Yooseph et al, 2010). The publicly available program PILER-CR (fitted with a custom post-filtering program to remove false-positives) was used to identify CRISPRs. CRISPRs were identified as short, highly conserved repeats separated by unique sequences of similar length and the analysis was performed on a Sanger-only assembly comprised of GOS Phase I and Indian Ocean data. The standing theory is that inter-repeat spacers represent samples of viruses and plasmids previously encountered by a given microbe. The spacer matches found in the GOS viral fraction were compared against the background matches found in microbial and viral subsets of the PANDA database to confirm that spacer sequences are indeed preferentially recruited from the viruses. A total of 144 CRISPR arrays were identified in this manner, with 31 of these arrays demonstrating spacer matches to viral metagenomic data generated from Indian Ocean samples. Eighteen scaffolds (12.5%) also contained CRISPR-associated (CAS) genes. CRISPR/Cas regions were primarily associated with Proteobacteria inhabiting hypersaline and coastal environments, with fewer originating from open-ocean locations. This observation correlated with the analysis of the Moore genomes, where CRSPR system was found primarily in microbes living in densely populated or spatially constrained environments such as bacterial mats or hydrothermal vents, as well as in hypersaline sites.

We are continuing to: (1) Examine the diversity and genomic context of viral cellular genes through phylogenetic and scaffold analyses, (2) Investigate the geographical distribution of viral cellular genes and use statistical analyses to determine if correlations exist between the abundance and distribution of certain gene families and environmental metadata (particularly nutrients), and (3) Determine if certain metabolic pathways are more enriched in lytic vs. temperate bacteriophages and eukaryotic viruses and evaluate how this may influence short-term vs. long-term adaptation of microbial hosts.

Task 3: Targeted coastal environments.

We have carried out a series of sampling efforts focused on the Southern California Bight and California Current Ecosystem in collaboration with several groups as discussed below.

CalCOFI: In the summer of 2007, we collaborated with the California Cooperative Oceanic Fisheries Investigations (CalCOFI) group to evaluate metagenomics at select areas characterized by coastal upwelling. These samples are a contrast to the primarily open ocean sites of the GOS expedition due to the upwelling of nutrient rich deep waters that enable an increase in primary production at the surface. The Southern California Bight is characterized by a unique water circulation pattern that results in seasonal upwelling, providing an opportunity to examine the impact of episodic pulses of cold nutrient rich water into surface communities. We encountered a large bloom of Planctomyces that comprises nearly 70% of the reads for a low diversity site within this ecosystem. We performed 454 and SOLID sequencing of this sample, providing an opportunity to potentially close the genome and conduct population genomics related analyses. Other analyses include:

1. *Correlations of sequence characteristics and nutrient availability.* Using APIS and COGs we found taxonomic differences among sites that correlate to nutrient concentrations. Sequences from 4 sites exhibit higher nutrients, Chl a and nitrate (upwelled) are dominated by CFB group (Cytophaga, Flavobacteria and Bacteriodetes), Gammaproteobacteria and Actinobacteria, whereas, 3 sites with

lower nutrient levels (oligotrophic), showed similar taxa to the GOS I primarily open ocean dataset (Rusch et al., 2007), which are typified by a dominance of Alphaproteobacteria and Cyanobacteria, primarily SAR11 and *Prochlorococcus*, respectively. Actinobacterial and Gammaproteobacterial sequences found in the upwelled-waters are most likely a novel clade as these sequences appear as deep branches among reference sequences. GC profiles and assembly analysis further confirm the link of composition to nutrient availability. A bimodal distribution of percent GC is seen in the eutrophic sites with a unique peak at ~ 50% GC. Site contributions to assembled scaffolds shows sequences from oligotrophic (same for eutrophic) sites assembling together at a statistically higher proportion than the co-assembly of oligotrophic and eutrophic sequences on scaffolds.

2. *Community composition of 0.8μm and 3.0μm filters.* Taxonomic profiles of bacterial sequences classified using APIS indicate increases in cyanobacteria specifically *Synechococcus* and *Prochlorococcus*, CFB group (known to be particle associated), and Gammaproteobacteria (nitrate utilizers). Also, the 3.0μm filter sequences show a marked increase of dsDNA viral sequences. Analysis of sequence similarity revealed that the viral populations on the 0.8 and 3.0μm filters exhibit clonal attributes. This finding supports our hypothesis that these sequences are indicative of a phytoplankton infection event, as viral particle numbers are much greater (~1000 per cell compared to ~50 per cell in bacteria).

3. *Annotation of biogeochemically relevant metabolic genes among viroplankton communities (viral metagenomes and larger size classes) and microbial communities to evaluate patterns (spatial) in the role of virioplankton and microbes on organic matter flux and food-web energetics.* Using APIS, we identified viral sequences from the larger filters and performed COG and HMM-based functional analyses on all sequences. Of particular significance is the finding of microbial nitrate/nitrite utilization genes within the viral metagenomes used in the assimilation of nitrate, as well as, homologs to nitric oxide reductase used in the dissimilation of nitrate to N_2 gas. Phylogenetics shows unique clading patterns for viral sequences compared to their bacterial counterparts. Additionally, evidence of viral-encoded beta-lactamases, antibiotic resistance genes, were identified on viral contigs post-assembly of reads; as well as, carbamyol phosphate synthase that is used in nucleotide and amino acid biosynthesis. Therefore, this dataset has evidence of carbon and nitrogen utilization genes within the viral communities, likely contributing to host fitness.

4. *Classification of micro-eukaryote sequences.* While this research is ongoing, initial evidence to understand which taxa are involved in utilizing the available nitrate in upwelled-waters (thus controlling the nitrogen cycle) indicates increases in green algae on the 0.8μm filter and Chromaveolates on the 3.0μm filter in the northern coastal sites. Future work will be directed to continued description of organisms present and identification of their metabolic potential to drive nitrogen and carbon cycling.

5. *Are viruses reservoirs of genes that can be used by microbes to gain access into new ecological niches?* To assess this long-standing hypothesis, analysis of sequences from virioplankton libraries is being used to facilitate correlations with the larger size fractions thus, elucidating the role of viruses and virus-mediated processes in the California Current microbial ecosystem.

Links between community function, genome content, and gene expression across gradients in light and carbon flow: In collaboration with the Scripps Institution of Oceanography, the focus of this work was to evaluate the changes in microbial communities over significant and experimentally characterized gradients in carbon fixation, nitrogen uptake, and Fe-light limitation. Over an 800 km transect from coastal California to the oligotrophic Pacific, samples were collected from 8 surface and subsurface chlorophyll maximum locations representative of coastal, mesotrophic, and oligotrophic environments. Extensive measurements conducted by multiple groups show that the metagenomic samples encompass over a 1000-fold gradient in total carbon flow in addition to dramatic changes in Fe and N utilization. Each filter fraction from the 8 metagenomic and metatranscriptomic samples has been sequenced using Sanger and Titanium 454. Additionally, metatranscriptomic libraries for

Eukaryotes and Prokaryotes were sequenced from each location. As nitrogen and carbon uptake was measured at the time of sampling, this dataset allows for a direct comparison between microbial gene expression and ecological function. A major highlight of this research is the dominance of Pelagophytes in nitrate uptake and thus carbon export at open ocean subsurface chlorophyll a maxima (**Figure 10**).

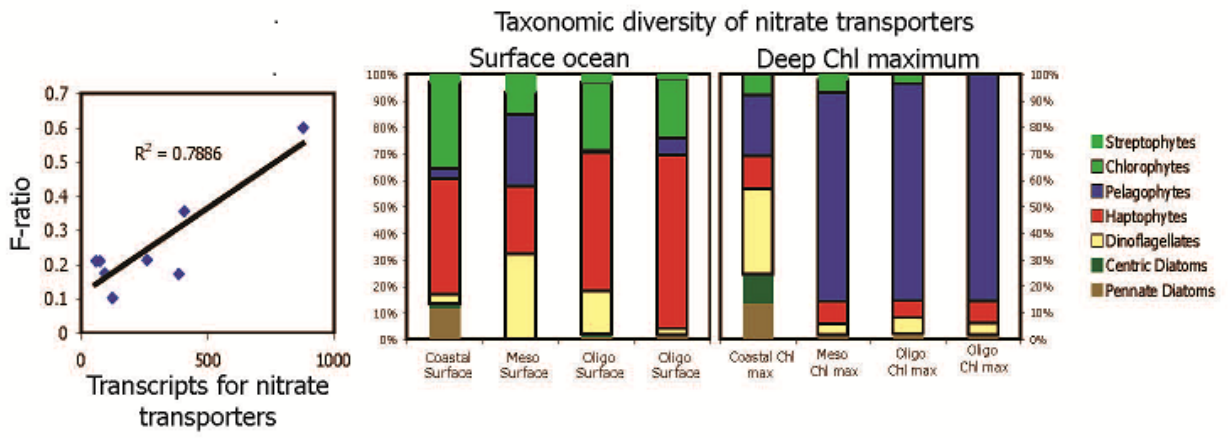


Figure 10. Eukaryotic gene expression and community function. On the left, the abundance of transporters for nitrate is compared to F-ratio, the proportion of total nitrogen uptake attributed to nitrate. The F-ratio is a proxy for carbon export from the surface ocean, and deep chl maxima are carbon export "hot spots." On the right the taxonomic affinity of the nitrate transports at each site are shown. The pelagophytes, best known for being the constituent of "brown tides" in Texas, appear to be responsible for greater than 75% of nitrate uptake in the deep chl maxima of the open ocean.

Monterey Bay: In collaboration with researchers at the Monterey Bay Research Aquarium (MBARI), the focus is to evaluate the interactive impacts of light and nitrate availability from onshore to offshore. Samples were collected from surface waters at an inner-shelf intense upwelling station (>20µM NO_3^-; >10ug/L chlorophyll), from surface and chlorophyll max communities at two mid-shelf stations of intermediate upwelling (5-10µM NO_3^-; 2-5ug/L chlorophyll), and from two off shore deep chlorophyll max stations.

Task 4: Additional Metagenomic Data Analysis Tool Development

We continue to develop and enhance metagenomic data analysis and visualization tools. Our goals are to streamline our high-throughput analysis pipelines, enhance and develop new tools and promote their public release and usability.

Development tools completed:

<u>Pipelines and Deployment of VICS</u>

One of the informatics infrastructure accomplishments was the deployment of VI Compute Server (VICS) which is a collection of grid-enabled bioinformatics tools with HTTP (browser) and Web Service (scriptable) interface (**Figure 11**). Its functionality allows simple execution of bioinformatics tools, as well as rapid implementation of stable pipelines for high throughput production. It provides a framework for scalable and robust implementations of annotation and data analysis pipelines. VICS is currently being used to run production metagenomic (prokaryotic and viral) pipelines at JCVI.

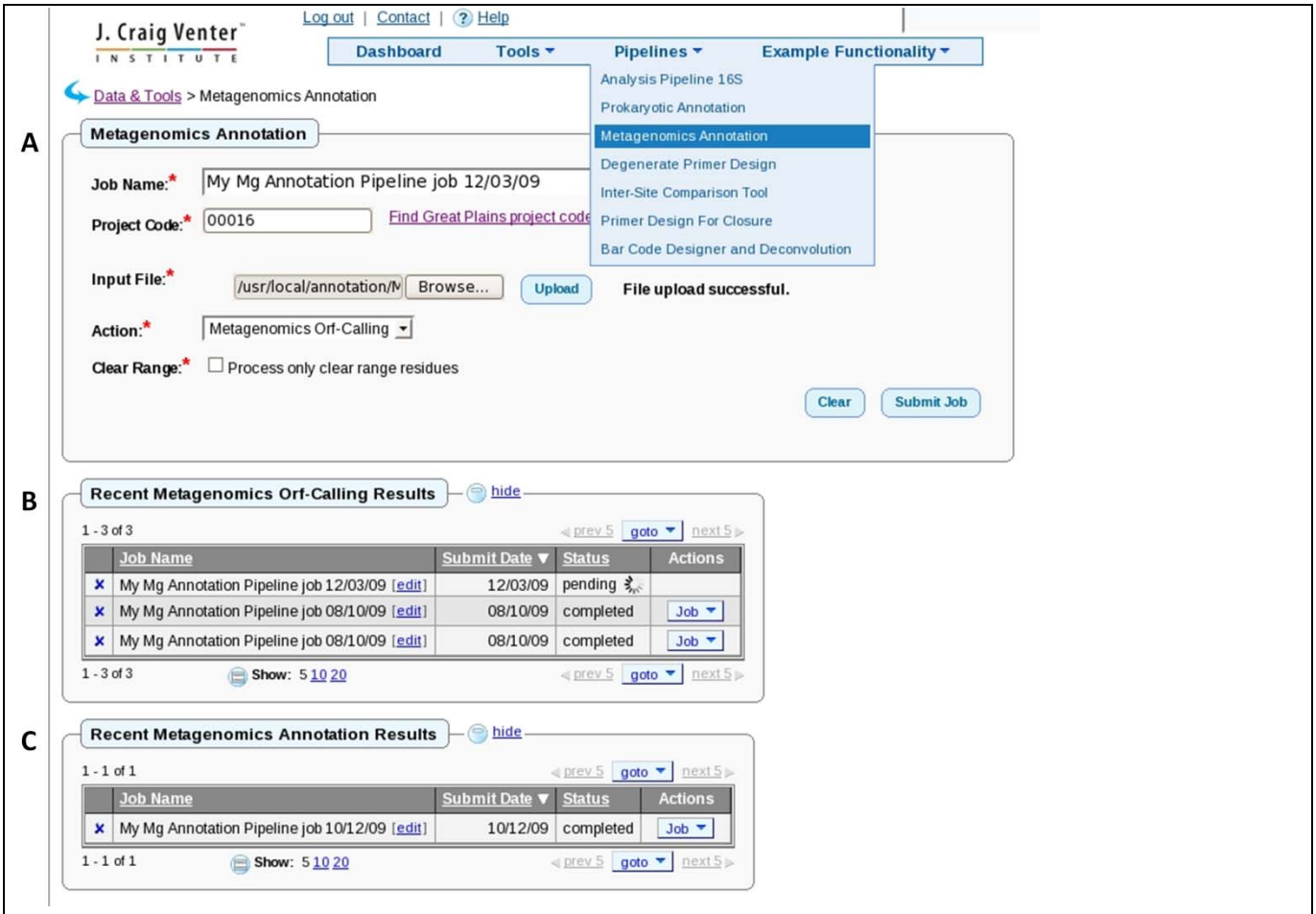


Figure 11. VICS Web Interface – The Metagenomics Annotation panel allows the user to specify a job name, select a valid JCVI project code, and specify a sequence input file (either nucleotide or peptide file) to be annotated (**A**). After clicking upload, the type of the file is automatically detected and the respective annotation mode is displayed (orf-calling for nucleotides, input annotation for amino acids). Here, the user has uploaded a nucleotide sequence file. The clear range option may be checked to indicate that only the clear range part of the sequences should be processed (e.g. for Sanger sequence data). After a job has been submitted, its execution status can be monitored for each pipeline type (**B & C**).

Prokaryotic metagenomic annotation pipeline

JCVI's metagenomic pipeline is based on its prokaryotic pipeline. The pipeline allows a choice of different metagenomic gene finders (METAGENE, FragGeneScan, and an in-house gene-finding method that tolerates frameshifts in the data) to do structural predictions. Functional annotation is classifying and attributing the identified structural elements by running several searches. The collections of attributes for each putative protein are then ranked and gene names are assigned using a hierarchy scheme. JCVI's compute resources has the capability of processing and annotating 1 million 454 reads in ~30 CPU hours using a dedicated grid resource. This pipeline can also identify and annotate genes on metagenomic assemblies.

Significant improvements were done to the JCVI's metagenomic prokaryotic pipeline this past year. This new pipeline uses the most up to date version of the Uniref 100 cluster database for the blastP searches on the predicted open reading frames. The HMM searches are done using the HMMer3 functionality against the most recent TIGRFAM and PFAM databases. The rules hierarchy that ranks the collection of attributes before the annotation summary is generated has been updated which has made significant improvements in the accuracy of Enzyme Commission (EC) & Gene Ontology (GO) assignments.

Viral metagenomic annotation pipeline

JCVI's viral metagenomic pipeline, developed in 2009, involves searches against several viral specific databases such as the ENV-nr and the ACLAME databases. The viral pipeline uses PhiGO which is a phage ontology associated with the ACLAME database for assigning GO terms. The search results are then loaded into a SQLlite database for easier querying and access. The summary annotation files with the gene name are then generated using a ranking system. JCVI's compute resources has the capability of processing and annotating 1 million 454 viral reads in ~50 CPU hours using a dedicated grid resource.

JCVI's metagenomic viral pipeline has been updated synchronizing it with the changes to the metagenomic prokaryotic pipeline. The Environmental databases have been updated with newer data and collapsed to reduce search time. This also uses the Uniref100 databases for blastP searches and the HMMer3 functionality for the HMM searches. The rules hierarchy has been updated and the output format has been changed to flat files.

Incremental clustering pipeline

The incremental clustering pipeline is a fast and reliable method of identifying and clustering new metagenomic proteins (Yooseph, Li et al. 2008). This pipeline is based on an efficient procedure that does not require the computationally expensive all against all computes. The protein clustering resource used by our various analyses is periodically updated using this pipeline.

Because of the large numbers of sequences being produced by next-generation sequencers, we streamlined the clustering pipeline by initially recruiting the incoming sequences against an RPS-blast database of the existing clusters. This replaces the initial cd-hit-2d clustering as well as a later blast recruitment stage. Novel reads are then clustered as before using cd-hit, and added to the database. This proved to be an efficient method to cluster the GOS circumnavigation 454 data.

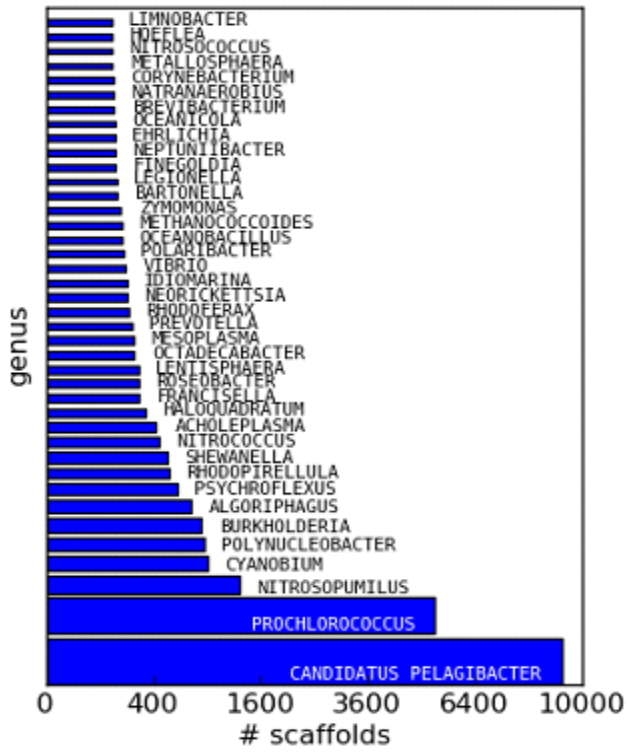
Fragment Recruitment

Fragment recruitment is a powerful graphical tool for comparing metagenomic reads directly to sequenced genomes to identify taxonomic composition, variability in gene content, and changes in genome structure. We have been revamping our fragment recruitment pipeline to improve its efficiency and by integrating it into our VICS services to provide improved accessibility and support public access to the code base. This process included implementing an automated mechanism to download and update our repository of microbial and viral genomes and their associated annotation. A number of changes were made to increase the speed of fragment recruitment. These changes include using a subset of the input reads to identify which genomes are likely to provide sufficient rates of recruitment to justify their searching against the entire dataset. We also implemented a mechanism where individual genomes are searched against only those genomes that are likely to generate interesting patterns of recruitment. This greatly reduces the time required to generate recruitment data and was particularly useful to recruit the 48 million GOS circumnavigation reads against the same set of reference genomes. Finally, we have automated the post-processing of the recruitment data and the transfer of this data to the server for visualization and analysis. In the future we look forward to modifying the recruitment pipeline to handle the massive volume of data generated by the Illumina and other next generation sequencers.

Other developments:

MGTAXA

In a continued development of our metagenomic classification pipeline MGTAXA, we added a methodology based on Glimmer Interpolated Context Models (ICMs) to make assignments of varying specificity to shorter contigs and unassembled 454 reads. Genus-level assignments for 5Kbp scaffolds in GOS metagenomic assembly are shown on **Figure 12**. The data collection, training and prediction are automated and parallelized to run on a cluster. This software is released under open source GNU license and can be obtained from the development site at http://andreyto.github.com/mgtaxa/. The method is also available as a public high-throughput computational Web server, where users can both run predictions against our pre-built models based on NCBI sequences and train their own custom models (http://mgtaxa.jcvi.org/).



APIS

APIS (Automated Phylogenetic Inference System) is an in-house tool used for phylogenetic and taxonomic classification and analysis of genomes. APIS was used to do taxonomic classification and phylogenetic analysis of the GOS circumnavigation data. APIS was successfully run on all GOS peptides. This resulted in a phylogenetic and functional assessment for all ORFs that had at least three homologs with a BLASTP evalue of 1e-9 or better. For samples sequenced by the traditional Sanger method, this was on average 62% of the peptides; for samples sequenced by pyrosequencing, 54%. We have also tied the results of APIS to transporter and other information in order to better utilize the data. This software will be made available as open source through sourceforge.

Figure 12. Distribution of scaffold counts assigned by MGTAXA at the level of genus. The area of each bar is proportional to the number of scaffolds assigned to a given genus. Genera with the top 40 counts are shown.

METAREP (Metagenomic Reports)

We developed a data viewing and analysis tool called METAREP (Goll et al. 2010) (**Figure 13**). It has a user friendly interface designed to help scientists view, analyze, and represent their data. It supports both functional and taxonomic browsing and offers flexible searching of annotated data using the Lucene search engine. This tool is being used to share GOS data with collaborators outside the institute in a secure manner and also as an infrastructure to display publicly available data. METAREP's comparative and visualization features have been greatly customized and enhanced recently to facilitate analysis of the GOS data. It currently comprises of a functional suite of tools for shotgun metagenomic data for viewing and analysis and provides graphical summaries for taxonomic and functional classifications, Gene Ontology, NCBI Taxonomy, and KEGG and MetaCyc Pathway Browser. Users can compare absolute and relative counts of multiple datasets at various functional and taxonomic levels. Advanced comparative features comprise statistical tests as well as multidimensional scaling, heat map and hierarchical clustering plots.

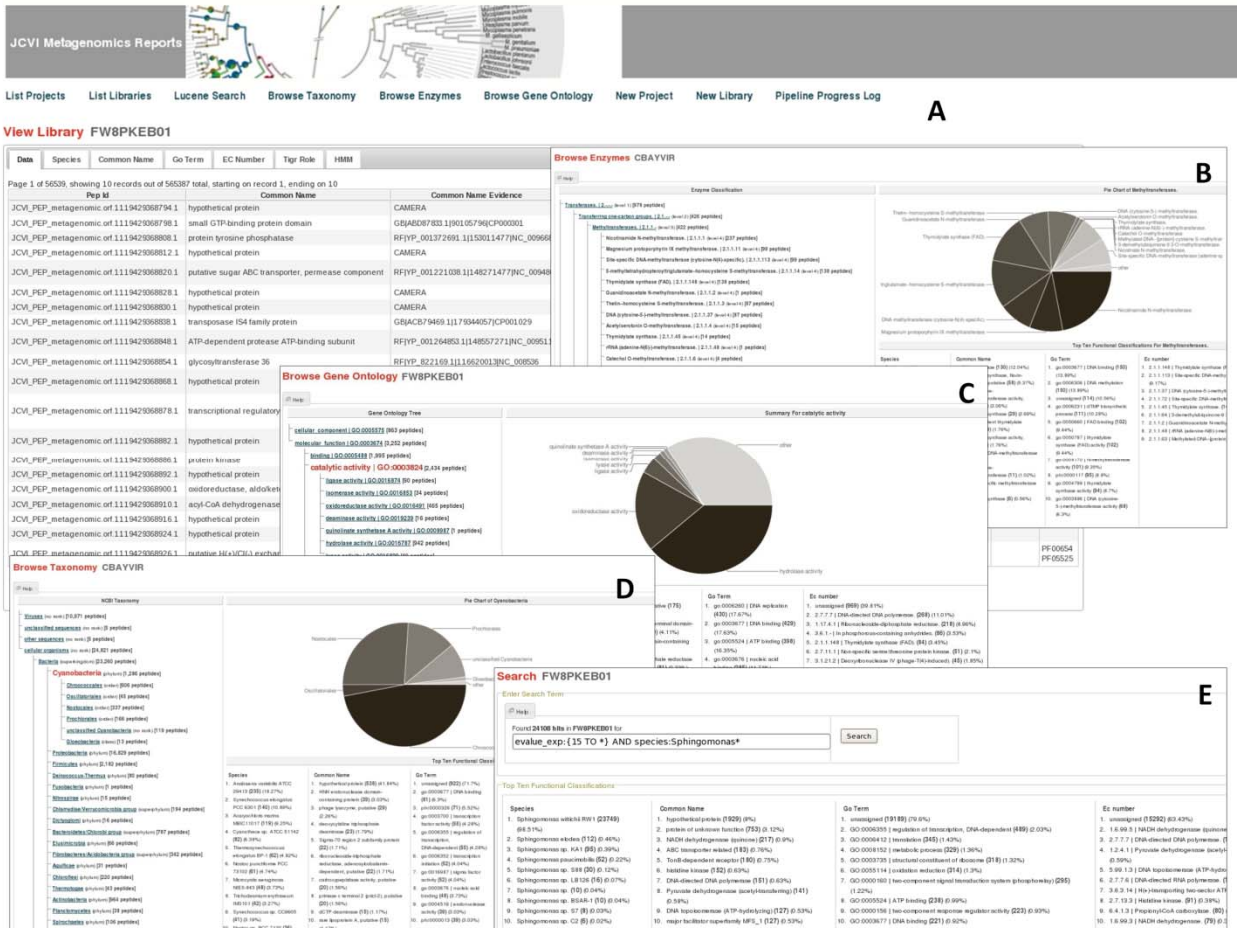


Figure 13. METAREP Web Interface –The METAREP library view page allows the user to browse the underlying annotation data, as well as to study the top 50 summaries of species, common name, Go, EC, TIGR Role Id and HMM assignments.(A). Peptide counts and annotation summaries can be studied using the METAREP Browse Enzymes (B), Browse Gene Ontology (C), and Browse Taxonomy (D) functionality. For each selected level, METAREP shows pie charts to summarize child counts and displays top ten lists for all other classification. To query and filter libraries, users can use the LUCENE query language which supports fielded data (E). Here, the user entered "evalue_exp:{15 TO * } AND Species:Sphingomonas*" to filter the data for peptides with a Blast E-value exponent of greater than 15 and a species assignment with genus *Sphingomonas* . Currently, 10 different fields are supported which can be logically combined using AND, OR, and NOT.

Tools for assessing community composition using markers

We developed high throughput analysis tools based on the methods we had for identifying core bacterial marker genes, estimating genome equivalents for normalization, and performing taxonomic profiling based on these core genes. A tool has also been built to visualize the normalized distribution of metagenomic samples across the tree of life and produce publication quality annotated large circular trees.

Transporter annotation

Membrane transporters play crucial role in fundamental cellular processes and functions in prokaryotic and eukaryotic genomes. For each organism, the complete set of membrane transport systems and outer membrane channels are predicted based on a series of bioinformatics evidence classified into different types and families according to their mode of transport, bioenergetics, molecular phylogeny and substrate specificity. We have a pipeline that does transporter annotation on whole genomes and has been tweaked to work on metagenomic data.

Metagenomic assembly

We have assembled all sequence data from the GOS circumnavigation expedition using the Celera Assembler software. This software is uniquely positioned to handle large volumes of Sanger+454 paired-end metagenomics data. The latest version of this software includes improved handling of multiple sequence alignments, including inconsistencies that typically result from the low-similarity thresholds used in this metagenomics assembly. JCVI successfully applied this same version to its Australian Soil metagenomics assemblies. The input to the GOS assembly was 20M Sanger reads and 28M 454 Titanium reads from 230 sites.

We have also developed a prototype pipeline for assembling and binning metagenomic scaffolds that correspond to closely related populations within a community. This pipeline utilizes a variety of signals including primary signals such as sequence overlap and mate pairing information and several secondary signals including oligonucleotide word frequencies, depth of coverage, sample distribution, and taxonomic assignment to arrive at the proper bins. We are currently testing and refining this pipeline and preparing a publication.

Task 5: Additional Sequencing.

As **Table 1** shows, in 2009 we utilized next-generation sequencing technology to more than triple the entire sequence read dataset of the Global Oceans Sampling project. The additional sequencing has added representation to the world's oceans of 16 previously un-analyzed GOS stations, explored additional communities by analyzing 128 previously un-analyzed samples, and increased the understanding of key datasets by adding and analyzing additional sequence data to 43 previously analyzed samples. Included in these samples are the larger size class of planktonic species which include the 0.8um-20um size range communities: picoplankton, large cyanobacteria, large phytoplankton and picozooplankton. We have sequenced nearly 50 new 0.8-3.0um size filters, nearly 50 new 3.0-20.0um size filters, and more than 15 new viral size filters. In addition we explored transcriptomics, a first for the project; and sequenced cDNA libraries for stations of the Scripps cruise in an effort to give a big-picture view of the microbial processes at different strata in the water column and in different ocean environments. No high throughput sequencing was done in 2010 but rather targeted additional sequencing of particular samples of interest. The final year of the award was dedicated to analysis of the data and publication of findings.

Table 1: GOS Sequencing Summary

STATIONS	**Stations**		
	A station is defined as one water filtering step in the field, usually resulting in 4 collected filter samples.		
	There are 241 GOS stations.		
		GOS cumulative total:	
		228 of 241 GOS stations have been analyzed.	
			Of the stations that have been analyzed:
			148 Stations were fully analyzed prior to 2009.
			16 Stations are newly represented on the globe in 2009.
			64 Stations received additional sequence data in 2009.
			228 Total stations have been analyzed.
SAMPLES	**Samples**		
	A sample is defined as one of the collected filters for a station.		
	Usually there are four filter samples collected for a station.		
	Each filter captures a different size range of micro-organism.		
	There are 933 samples collected for all GOS, thus far.		
	Not all samples have been analyzed.		
		GOS cumulative totals:	
		220	0.1µm filter samples have been analyzed.
		59	0.8µm filter samples have been analyzed.
		56	3.0µm filter samples have been analyzed.
		22	viral filter samples have been analyzed.
		357	Total samples have been analyzed.
			Prior to 2009:
			172 0.1µm filter samples were analyzed.
			9 0.8µm filter samples were analyzed.
			5 3.0µm filter samples were analyzed.
			0 viral filter samples were analyzed.
			186 Total samples were analyzed prior to 2009.
			In 2009, this many filter samples are analyzed for the first time:
			13 0.1µm filter samples.
			49 0.8µm filter samples.
			49 3.0µm filter samples.
			17 viral filter samples.
			128 Total filter samples are analyzed for the first time.
			In 2009, additional sequence data are generated for this many filter samples:
			35 0.1µm filter samples.
			1 0.8µm filter samples.
			2 3.0µm filter samples.
			5 viral filter samples.
			43 Total filter samples where additional data are generated.
SEQUENCE DATA	**Amount of sequence data**		
	Sequence data can be defined in terms of DNA sequence reads and total base pairs.		
	Sanger sequencing data were only generated prior to 2009.		
	454-Titanium (pyro-sequencing) sequence data were generated in 2009.		
		Sanger sequence totals (generated prior to 2009):	
		18,918,696	Sanger sequence reads for all GOS.
		14,126,590,303	Sanger base pairs for all GOS.
		454-Titanium sequence totals (generated in 2009):	
		65,332,726	454 sequence reads for all GOS (estimated through February 2010).
		22,619,771,385	454 base pairs for all GOS (estimated through February 2010).
	GRAND TOTALS: Sanger plus 454 data:		
		84,251,422	Sanger and 454 **reads** combined (estimated through February 2010).
		36,746,361,688	Sanger and 454 **base pairs** combined (estimated through February 2010).

GOS Publications

1. Rusch, D.B., et al., **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific.** PLoS Biol, 2007. **5**(3): p. e77.
2. Yooseph, S., et al., **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families.** PLoS Biol, 2007. **5**(3): p. e16.
3. Kannan N., Taylor S.S., et al., **Structural and functional diversity of the microbial kinome.** PLoS Biol, 2007. 5(3).
4. Seshadri, R., et al., **CAMERA: a community resource for metagenomics.** PLoS Biol, 2007. **5**(3): p. e75.
5. Yooseph, S., W. Li, and G. Sutton, **Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering.** BMC Bioinformatics, 2008. **9**(1): p. 182.
6. Williamson, S.J., et al., **The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples.** PLoS ONE, 2008. **3**(1): p. e1456.
7. Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M, Johnson J, Montgomery R, Ferriera S, Beeson K, Williamson SJ, Tovchigrechko A, Allen AE, Zeigler LA, Sutton G, Eisenstadt E, Rogers YH, Friedman R, Frazier M, Venter JC. **Genomic and functional adaptation in surface ocean planktonic prokaryotes.** *Nature*. 2010 Nov 4;468(7320):60-6.
8. Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. **Characterization of Prochlorococcus clades from iron-depleted oceanic regions.** *Proc Natl Acad Sci U S A.* 2010 Sep 14;107(37):16184-9.
9. Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, Woyke T, Welsh RM, Ishoey T, Lee JH, Binder BJ, DuPont CL, Latasa M, Guigand C, Buck KR, Hilton J, Thiagarajan M, Caler E, Read B, Lasken RS, Chavez FP, Worden AZ. **Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton.** *Proc Natl Acad Sci U S A.* 2010 Aug 17;107(33):14679-84. Epub 2010 Jul 28.
10. Tanenbaum DM, Goll J, Murphy, S, Kumar P, Zafar N, Thiagarajan M, Madupu R, Davidsen T, Kagan L, Kravitz S, Rusch DB, Yooseph S. **The JCVI Standard Operating Procedure for Annotating Prokaryotic Metagenomic Shotgun Sequencing Data.** Stand. Genomic Sci. 2010 Mar 25; 2(2).
11. Thomas T, Rusch D, Demaere MZ, Yung PY, Lewis M, Halpern A, Heidelberg KB, Egan S, Steinberg PD, Kjelleberg S. **Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis.** *ISME* J. Jun 3. 2010.
12. Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methé BA, Yooseph S. **METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics.** *Bioinformatics*. 2010 Oct 15;26(20):2631-2. Epub 2010 Aug 26.
13. Gilbert JA, Dupont CL. **Microbial Metagenomics: Beyond the Genome.** *Ann Rev Mar Sci.* 2011;3:347-71.
14. Allen LZ, Ishoey T, Novotny MA, McLean JS, Lasken RS, Williamson SJ. **Single virus genomics: a new tool for virus discovery.** *PLoS One.* 2011 Mar 23;6(3):e17722.
15. Wu D., Wu M., Halpern A., Rusch D.B., Yooseph S., Frazier M., Venter J.C., Eisen J.A. **Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees.** *PLoS One.* 2011;6(3).
16. Lorenzi, H. A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L., Williamson, S. J. **The Viral MetaGenome Annotation Pipeline (VMGAP): An Automated Tool for the Functional Annotation of Viral Metagenomic Shotgun Sequencing Data.** *Standards in Genomic Sciences.* 2011; 4(3).
17. Worden AZ, Dupont, C., Allen A. E. **Genomes of uncultured eukaryotes: sorting FACS from fiction**. Genome Biology 2011; 12:117.

18. Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J., Richter, R. A., Valas, R., Novotny, M. Yee-Greenbaum, J., Selengut, J. D., Haft, D. H., Halpern, A. L., Lasken, R. S., Nealson, K., Friedman, R., Venter, J. C. **Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage.** *ISME J.* online early 2011.

19. Allen, LZ, Allen, E.E., Badger, J. H., McCrow, J. P., Paulsen, I. T., Elbourne, L. D. H., Thiagarajan, M., Rusch, D. B, Nealson, K. H., Williamson, S. J., Venter, J. C., Allen, A. E. Influence of nutrients and currents on the genomic composition of microbes across an upwelling gradient. *ISME J* online early 2012

In Preparation

1. The genome encyclopedia for surface marine bacteria, Yooseph, Dupont, Nealson, Rusch et al.

2. Pelagophyceae contribute disproportionately to carbon fixation and nitrate in the open ocean. Chris L. Dupont, Ruben Valas, Ahmed Mustafa, John P. McCrow, Andrew E. Allen (J. Craig Venter Institute); Katherine Barbeau, Brian Palenik, Rhona Stuart, Amanda Chan, Randelle Bundy, Kelly Roe (Scripps Institution of Oceanograpy, UCSD); Zackary Johnson (Duke University); Elizabeth Mann (Skidaway Institution of Oceanography) *(ISME J)*

3. Metagenomic Exploration of Viruses throughout the Indian Ocean. Williamson, S. J; Allen, L. Z.; Lorenzi, H. A.; Fadrosh, D. W.; Kattnig, H.; Badger, J.; McCrow, J.P.; Tovchigretchko, A., Yooseph, S. et al. (*PLoS One)*

4. Groundtruthing Next-Gen Microbial Ecology-Biases and Errors in Community Structure Estimates Obtained from 454 Titanium PCR Amplicon Sequencing. Herbold, C.W.; Lee, C.K.; Polson, S.W.; Wommack, K.E.; Williamson, S.J.; Cary, S.C. (*ISME Journal)*

5. Influence of habitat on ecology and genome evolution of surface ocean bacterioplankton. Andrew E. Allen, John P. McCrow, Christopher L. Dupont, Douglas B. Rusch, Jonathan H. Badger et al. (*Science)*

6. Diversity and abundance of major classes of eukaryotic peptides in the surface ocean. Ruben Valas, Chris L. Dupont, John P. McCrow, Jonathan H. Badger, Douglas B. Rusch, Andrew E. Allen (*Genome Research/Genome Biology)*

7. Large scale patterns of gene expression in a Puget Sound Phytoplankton Bloom. Ahmed, Andrew E. Allen. (*PLoS Genetics)*

8. Transcript profiles of nitrogen and iron limited phytoplankton in simulated upwelling mesocosm. Andrew E. Allen, John P. McCrow, Badger, Chris L. Dupont, collaborators (*PLoS Biology)*

Goal 2: Synthesize a minimal mycoplasma genome that has all of the machinery for independent life.

The following specific aims were presented in our 2005 proposal to DOE:

1. Construct a natural *Mycoplasma genitalium* chromosome from synthetic oligonucleotides.
2. Transplant the chromosome into a cell from which the natural genome has been removed or destroyed.
3. Define a set of genes that are collectively dispensable by using acridine orange to create multiple frameshifts.
4. Construct a reduced or minimal genome by removing dispensable genes.
5. Construct an *in vitro* homologous recombination system from *Deinococcus radiodurans* that can efficiently assemble numerous overlapping synthetic DNA pieces.

With the goal of synthesizing a minimal bacterial genome that will facilitate the investigation of the essential machinery for independent life, in 2010 the JCVI Synthetic Biology team announced the creation of a *Mycoplasma mycoides* cell with a fully synthetic genome (8). To achieve this milestone along our path to build a bacterial cell with a genome encoding only the minimal number of genes necessary for life, we developed a suite of new tools that will advance the whole field of synthetic biology. Many of these tools were different than originally proposed (e.g., assembly was not accomplished using the *in vitro* homologous recombination system from *Deinococcus radiodurans*).

We used these molecular tools to synthesize two near minimal bacterial genomes and to boot up one of those genomes using a technique called genome transplantation to make what is popularly called the "synthetic cell". These accomplishments were seminal steps towards our aim of discovering fundamental new knowledge about requirements for microbial life and laying a basic research foundation for developing microbiological approaches to bioenergy.

Note that whereas DOE provided the earliest support for our work in this area beginning in 2003, we were also able to obtain additional funds from Synthetic Genomics, Inc. (SGI) to substantially grow synthetic genomics activities at JCVI. DOE supported our minimal genome research between 2003 and September 2005, and then again from July 2010 through August 2011. Between September 2005 and July 2010, the program was funded exclusively by SGI.

Specific Aims. Our goal in this aspect of the program is to create a minimal bacterial cell based on a group of bacteria called mycoplasmas. These organisms have the smallest genomes of any bacterial cells that can be grown in pure culture. A minimal cell would contain only the essential genes needed to be grown in pure culture under defined laboratory conditions. It would lack synthetic capacity for small molecules or metabolites that can be supplied in the medium. Thus it would be stripped down to core functions for macromolecular synthesis and cell division. The rational for this is that through creation and analysis of a cell with perhaps fewer than 400 protein coding genes we will be better able to learn the first principals of cellular life. Such a cell would have less than one tenth as many genes as *Escherichia coli* and the lack of complexity would enable an uncluttered perspective on how cells work.

Background. Mycoplasmas have the smallest genomes of any organisms that can be grown in pure culture. More than a dozen years ago, we sequenced *M. genitalium* strain G37 (funded by DOE) and annotated 485 protein-coding genes and 43 RNA genes. *M. genitalium* has the smallest genome of among the known mycoplasmas. While *M. genitalium* was believed to be nearly a minimal cell in nature, it carried more genes than necessary for growth under ideal conditions in the laboratory. We carried out global transposon mutagenesis in two separate studies, the first in 1999 and the second in 2006 (11, 12). These studies identified

115 protein-coding genes that could be disrupted one-at-a-time without loss of viability, although some of these knockouts altered the phenotype in such a way that they may not be entirely dispensable. It remains an unanswered question as to how many of the 115 genes are simultaneously removable without loss of viability. Because mycoplasmas have almost no genomic redundancy, there should be very cellular functions encoded by more than one gene such that at least one gene must be kept to perform an essential function. Still cumulative effects from the removal of multiple genes may lower the fitness of the deletion mutant to the point that growth is too slow to be observable.

As one of our specific aims in this process we sought to create an *M. genitalium* strain with a large number of genes simultaneously disrupted by repeatedly passaging the organism in media containing the frameshift inducing mutagen acridine orange. We did this, and after 20 iterative passages (more than one year of culture) produced an *M. genitalium* strain that grew much more slowly than the wild type. Sequence analysis of that mutant showed there were hundreds of substitution mutations and only a handful of frameshift mutations. This approach to producing a cell with multiple disrupted genes was abandoned because the substitution mutations made interpretation of the experiment impossible.

Manipulating Mycoplasma Genomes. When we began this project there were and remain few genetic tools for manipulating mycoplasma genomes. While we could readily disrupt a single gene in a genome for some species of mycoplasmas, there were no methods to remove sequentially many different genes to get down to the essential gene set for laboratory growth. This led us to the strategy of building a synthetic mycoplasma cell encoding only a minimal gene set. The effort to create these new methods of genome assembly and genome transplantation, which are absolutely necessary to achieve the goals of our DOE minimal cell project, was beyond the budget of the DOE grant. Instead this work has been funded by a grant to the JCVI from Synthetic Genomics, Inc. (SGI). SGI sought these methods to create synthetic and semi-synthetic cells for use in materials production, such as biofuels.

Towards the goal of manipulating mycoplasma genomes we have developed new enabling technologies for the field of Synthetic Biology. Most importantly, we can now isolate a bacterial genome from one cell; park that genome in a eukaryotic vector species, the yeast *Saccharomyces cerevisiae*; then isolate that genome from its yeast vector, and transplant it into a suitable bacterial recipient cell (16). Importantly for our minimal cell project, once the genome is in yeast it can be genetically manipulated using the enormous power of yeast genetic tools (16). We have also developed new methods for mycoplasma biology and new approaches for cloning bacterial genomes as yeast artificial chromosomes (1-3, 14, 17).

While our original aim was to create a minimal cell based on the synthetic *M. genitalium* genome we reported in 2008 (7), we still have not managed to boot up that genome via genome transplantation. Because of the exceedingly slow growth of this species that slowed the pace of experimentation and a potent set of extracellular nucleases encoded by our target recipient cell for these experiments that degraded our donor genomes, we adopted a different set of mycoplasmas as a research platform. Our choice was *M. mycoides*, which is a goat pathogen with a genome almost twice the size of the *M. genitalium* genome and that grows ~10 times faster than *M. genitalium*. As reported in our previous progress report, because we already knew how to boot up an *M. mycoides* genome by transplanting it into an *M. capricolum* cell (15, 16), we elected to synthesize a 1.1 Mbp *M. mycoides* genome. The plan was first to make a synthetic cell that had almost the entire *M. mycoides* gene set. This would establish the technology using a genome we had every reason to expect to we could boot up using genome transplantation. Then later we would begin wholesale genome reduction to create a minimal cell. In the July 2nd 2010 issue of **Science** we reported our synthetic *M. mycoides* cell. So that the synthetic *M. mycoides* genome could be distinguished from a natural one, we included a number of watermarks in the sequence and either removed or inactivated 13 other genes (including gene deletions we expected would make the organism non-pathogenic to goats) (8).

Once we completed the creation of a synthetic mycoplasma cell, SGI funded efforts on mycoplasmas ceased. Since that point in the middle of 2010 the JCVI has been refining tools that created the synthetic cell to create new versions of *M. mycoides* that contain fewer and fewer genes.

Strategies for *M. mycoides* Genome Minimization. While the bulk of our progress towards the foundational technology needed to create a minimal cell has been funded by SGI, we have used DOE funds to develop ancillary technologies and data sets that will accelerate minimal cell generation once we have a synthetic cell in hand. At the outset of our project there were essentially no genetic tools for *M. mycoides*. We have enabled robust homologous recombination-based mutagenesis of *M. mycoides* by expressing *E. coli* RecA protein via a suicide plasmid in *M. mycoides*. This yields single gene knockouts much more rapidly than we can do using our yeast cloning approach (2). We have also developed a synthetic growth media for *M. genitalium* that will enable us to perform more instructive metabolism experiments with *M. genitalium* based minimal cells (19). A version of this synthetic growth media also works well for *M. mycoides*. Earlier in this DOE funded project we developed a single gene disruption approach for mycoplasmas using TN4001 transposon bombardment to identify non-essential genes in *M. genitalium (11, 12)*. In that method development we learned how to create single gene disruption mutants in *M. capricolum* and *M. mycoides,* which are fast growing relatives of *M. genitalium*.

To determine what genes we can remove from the *M. mycoides* genome, we employed both comparative genomics and global transposon mutagenesis strategies. We identified the genes conserved in a set of 14 different mycoplasma species (including 7 that we sequenced). We compared the non-essential gene sets from the DOE supported *M. genitalium* (11, 12) studies with the sets from two other exhaustive essential gene studies of related mycoplasmas (5) (Glass unpublished results). We also used data from an unpublished study in 2005 where we isolated transposon insertion mutants for 201 non-essential *M. capricolum* genes. Our alternate platform organism for minimal cell construction, *M. mycoides* is 91.5% identical at the nucleotide level to *M. capricolum*. Not unexpectedly our analysis showed that orthologs of non-essential genes in *M. genitalium* were also not essential in *M. capricolum, M. pulmonis* and *M. alligatoris*. This gave us one prediction of non-essential genes in *M. mycoides*. We also used 454 pyrosequencing to identify the transposon insertion sites for a large population of *M. mycoides* mutants. That identified 675 genes with TN4001 insertions of the 917 *M. mycoides* genes (**Figure 13**). Because the *M. mycoides* genome has many more paralogous gene families than *M. genitalium,* we believe the essential gene set for this organisms will contain considerably more than the 242 protein coding genes that did not have transposon insertions. Our plan for excision of genes will be based on both the comparative genomics and transposon studies.

Current and Future Work to Construct a Minimal Cell. Our research has developed methods for systematic removal of non-essential genes from *M. mycoides*. We can iteratively remove genes from the *M. mycoides* genome cloned as a YAC and then transplant the reduced genome into *M. capricolum* cells to assess whether the deleted gene was non-essential and the transplantation yielded viable cells, or if the no transplants were obtained, that the gene was likely essential. However with a 2 week cycle between gene removals, this approach would likely take years to complete. Accordingly, we devised methods for greatly accelerating the minimization of the genome that enable us to remove multiple genes simultaneously that both employ the methods we developed for DNA synthesis and assembly (6, 9, 10). Key to our strategy is the fact that we have 11 overlapping sub-genomic fragments that we can transform into yeast to assemble the entire synthetic *M. mycoides* genome (8). These 11 sub-genomic fragments are intermediates we created during the synthesis of the *M. mycoides* genome. They are all confirmed to be functional. This means that we can begin deleting genes from each sub-genomic fragment in parallel to produce pool of reduced genome fragments. These can then be combinatorially reassembled in yeast to yield many different genomes. The genomes can then be tested for functionality by performing genome transplantation and looking for viable cells. For gene deletion we are currently evaluating two different strategies. Both rely on engineering of complete bacterial genomes harbored in yeast. One of these is the TREC method (18), which we developed for making seamless deletions of any genomic regions that is cloned in yeast. More recently we have started to apply the new "Green Monster" technology (20) to make multiple simultaneous deletions in the synthetic genome while it is parked in a yeast vector as a YAC.

Figure 13

Mycoplasma mycoides JCVI-syn1.0

TREC, which stands for Tandem Repeat Coupled with Endonuclease Cleavage, was developed to manipulate bacterial genomes cloned in yeast because existing methods led to instability of these genomes (18). In this method, a DNA cassette used to make a genomic change is precisely and efficiently removed when downstream portion of the cassette designed to recombine with the genomic region upstream of the cassette is induced to do so (via generation of free ends by tools within the cassette). We are now using our knowledge of non-essential M. mycoides genes to make deletions in multiple sub-genomic segments of the bacterial genome using TREC. Directly pursuant of our plan to independently remove non-essential genes from each of the 11 sub-genomic fragments, we have systematically deleted 36 non-essential genes from one of the 11 fragments. We have also identified other genes in that sub-genomic fragment that were essential or were part of a paralogous gene family encoding essential functions. Additionally, in other sub-genomic fragments, we have used this strategy to remove all 6 sets of restriction modification genes from the genome, as well as a set of 3 of genes involved in glycerol metabolism. The elimination of the restriction enzymes improves our efficiency of transplantation, perhaps because the unmethylated genome transplanted from yeast was degraded by enzymes encoded by the donor genome. The elimination of the glycerol metabolism genes likely rendered the mutants non-pathogenic. The mutants no longer produce detectable H2O2 and hemolyze red blood cells (manuscript in preparation). Although M. mycoides is a veterinary rather than a human pathogen, we were glad to have made the synthetic cell avirulent. We have also used TREC to eliminate the ~30 Kbp integrative conjugal element. In sum these TREC modifications produced genomes lacking about 10% of the genes present in wild type M. mycoides.

Green Monster technology has also been used to create a genome lacking a different set of genes. The developer of this technology for generating multi-deletion genomes (20) recently joined our group. In this method, deleted regions are replaced by a with a green fluorescent protein (GFP) reporter gene. Mutants with GFP genes replacing different *M. mycoides* genes are rapidly assembled into a single genome via repeated rounds of mating, meiosis, and flow cytometry-based enrichment of the host yeast cells that with the greatest fluorescence. To date, we have developed an improved GFP deletion cassette, tools for linearizing and re-circularizing the *M. mycoides* genomes, and a method to move the bacterial genomes to a compatible yeast genetic background. At the conclusion of the DOE funding on this project we had completed 3 cycles of gene excision from the *M. mycoides* yeast clone to produce a genome lacking 8 clusters of genes we believe to be non-essential. Transplantation of that genome yielded an *M. mycoides* lacking ~10% of its original gene complement and demonstrated the effectiveness of the Green Monster approach.

By the time we concluded work on this project in late summer 2011, both the TREC and Green Monster approaches had effectively been used to generate versions of the *M. mycoides* synthetic cell lacking about 10% of the wild type genome. The Green Monster approach was faster than using TREC; however it will not work for producing fully minimized genome. While TREC results in clean deletions, each Green Monster deletion leaves a GFP gene at the insertion site. Thus a fully minimized genome might have between 100 and 200 GFP genes. So many GFP genes would exceed the capacity of flow cytometry to identify new GFP replacements of deleted genes in the yeast cloned genome based on increased fluorescence. Additionally, the effect of so many GFP genes in the resulting *M. mycoides* cells could confound the utility of a cell minimized by the Green Monster approach for basic cell function studies. Still Green Monster is effective for rapidly guiding a TREC deletion strategy. At the conclusion of this DOE project we had developed a combined TREC-Green Monster approach, and had a significantly reduced *M. mycoides* cell that was both non-pathogenic and devoid of restriction endonucleases, which could complicate future work. At that point we set aside our minimal cell work with the intent of starting again once we had obtained new funding for the effort. This organism and the effort to evaluate these two genome minimization strategies will be the basis of our future studies on producing and analyzing a minimal bacterial cell. Furthermore, we anticipate that both the TREC and Green Monster approaches will have wide applicability beyond *M. mycoides.*

It is an appropriate note to add to this report that in 2012 both SGI and DARPA elected to continue funding the JCVI efforts to produce a minimal bacterial cell. Both funders saw the value in creating such a cell to use as a platform for understanding the first principles of cellular life.

Spinoff Accomplishments Resulting from the minimal genome project.

Medical Microbiology. Even though our efforts to develop a minimal bacterial cell have the ambition of elucidating the first principles of cellular life rather than addressing issues in mycoplasma pathogenesis, Reagents and techniques developed using DOE funding have resulted in several advances in medicine. Transposon insertion mutants in *M. capricolum* were critical in a study done at the University of Florida showing a method for *M. mycoides* site directed mutagenesis that could help in the eventual development of therapies for a close relative of the *M. mycoides* strains we worked with that cause a cattle disease endemic in equatorial Africa called contagious bovine pleuro pneumonia (CBPP) (2). There is currently no effective vaccine for CBPP. The synthetic genomics methods we developed that resulted in the synthetic *M. mycoides* cell are now being used at the JCVI in a NSF-Gates Foundation funded project to develop a live attenuated *M. mycoides* to prevent CBPP.

The set of 115 different *M. genitalium* mutants developed during the DOE funded phases of this project have been shared with many research groups investigating this sexually transmitted human pathogen. Graduate students at the University of Washington and the University of Texas Health Science Center at San Antonio used the strains as part of their dissertation projects. To date this has resulted in work showing genes responsible for *M. genitalium* evasion of the human immune system (4). In a JCVI collaboration with Scripps Research Institute, *M. genitalium* transposon mutants have been used to produce the first evidence that *M. genitalium* infection could trigger multiple myeloma, which is one of the most common human lymphomas. The work even suggests possible strategies for prevention of most cases of this incurable cancer.

Systems Biology & Minimal Cell Studies. The *M. genitalium* mutants and science generated from our study of this minimal bacterium inspired a landmark work in systems biology. Markus Covert and his team at Stanford University in collaboration with JCVI developed a computer simulation of *M. genitalium* that takes into account all known facts about the biology of this simple organism. It was used to predict the phenotypes of a number of *M. genitalium* mutations. Those predictions were compared to the characteristics of several of the JCVI *M. genitalium* mutants (13).

References for Goal 2

1. **Algire, M. A., C. Lartigue, D. W. Thomas, N. Assad-Garcia, J. I. Glass, and C. Merryman.** 2009. New selectable marker for manipulating the simple genomes of Mycoplasma species. Antimicrob Agents Chemother **53:**4429-4432.
2. **Allam, A. B., L. Reyes, N. Assad-Garcia, J. I. Glass, and M. B. Brown.** 2010. Enhancement of targeted homologous recombination in Mycoplasma mycoides subsp. capri by inclusion of heterologous recA. Appl Environ Microbiol **76:**6951-6954.
3. **Benders, G. A., V. N. Noskov, E. A. Denisova, C. Lartigue, D. G. Gibson, N. Assad-Garcia, R. Y. Chuang, W. Carrera, M. Moodie, M. A. Algire, Q. Phan, N. Alperovich, S. Vashee, C. Merryman, J. C. Venter, H. O. Smith, J. I. Glass, and C. A. Hutchison, 3rd.** 2010. Cloning whole bacterial genomes in yeast. Nucleic Acids Res **38:**2558-2569.
4. **Burgos, R., G. E. Wood, L. Young, J. I. Glass, and P. A. Totten.** 2012. RecA mediates MgpB and MgpC phase and antigenic variation in Mycoplasma genitalium, but plays a minor role in DNA repair. Infect Immun **In Press**.
5. **French, C. T., P. Lao, A. E. Loraine, B. T. Matthews, H. Yu, and K. Dybvig.** 2008. Large-scale transposon mutagenesis of Mycoplasma pulmonis. Mol Microbiol **69:**67-76.

6. **Gibson, D. G.** 2009. Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. Nucleic Acids Res **37:**6984-6990.
7. **Gibson, D. G., G. A. Benders, C. Andrews-Pfannkoch, E. A. Denisova, H. Baden-Tillson, J. Zaveri, T. B. Stockwell, A. Brownley, D. W. Thomas, M. A. Algire, C. Merryman, L. Young, V. N. Noskov, J. I. Glass, J. C. Venter, C. A. Hutchison, 3rd, and H. O. Smith.** 2008. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science **319:**1215-1220.
8. **Gibson, D. G., J. I. Glass, C. Lartigue, V. N. Noskov, R. Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z. Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison, 3rd, H. O. Smith, and J. C. Venter.** 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. Science **329:**52-56.
9. **Gibson, D. G., H. O. Smith, C. A. Hutchison, 3rd, J. C. Venter, and C. Merryman.** 2010. Chemical synthesis of the mouse mitochondrial genome. Nat Methods **7:**901-903.
10. **Gibson, D. G., L. Young, R. Y. Chuang, J. C. Venter, C. A. Hutchison, 3rd, and H. O. Smith.** 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods **6:**343-345.
11. **Glass, J. I., N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, 3rd, H. O. Smith, and J. C. Venter.** 2006. Essential genes of a minimal bacterium. Proc Natl Acad Sci U S A **103:**425-430.
12. **Hutchison, C. A., S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, and J. C. Venter.** 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. Science **286:**2165-2169.
13. **Karr, J. R., J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival, N. Assad-Garcia, J. I. Glass, and M. W. Covert.** 2012. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell **In Press**.
14. **Krishnakumar, R., N. Assad-Garcia, G. A. Benders, Q. Phan, M. G. Montague, and J. I. Glass.** 2010. Targeted chromosomal knockouts in Mycoplasma pneumoniae. Appl Environ Microbiol **76:**5297-5299.
15. **Lartigue, C., J. I. Glass, N. Alperovich, R. Pieper, P. P. Parmar, C. A. Hutchison, 3rd, H. O. Smith, and J. C. Venter.** 2007. Genome transplantation in bacteria: changing one species to another. Science **317:**632-638.
16. **Lartigue, C., S. Vashee, M. A. Algire, R. Y. Chuang, G. A. Benders, L. Ma, V. N. Noskov, E. A. Denisova, D. G. Gibson, N. Assad-Garcia, N. Alperovich, D. W. Thomas, C. Merryman, C. A. Hutchison, 3rd, H. O. Smith, J. C. Venter, and J. I. Glass.** 2009. Creating bacterial strains from genomes that have been cloned and engineered in yeast. Science **325:**1693-1696.
17. **Noskov, V. N., R. Y. Chuang, D. G. Gibson, S. H. Leem, V. Larionov, and N. Kouprina.** 2011. Isolation of circular yeast artificial chromosomes for synthetic biology and functional genomics studies. Nat Protoc **6:**89-96.
18. **Noskov, V. N., T. H. Segall-Shapiro, and R. Y. Chuang.** 2010. Tandem repeat coupled with endonuclease cleavage (TREC): a seamless modification tool for genome engineering in yeast. Nucleic Acids Res **38:**2570-2576.
19. **Suthers, P. F., M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas.** 2009. A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. PLoS computational biology **5:**e1000285.
20. **Suzuki, Y., R. P. Onge, R. Mani, O. D. King, A. Heilbut, V. M. Labunskyy, W. Chen, L. Pham, L. V. Zhang, A. H. Tong, C. Nislow, G. Giaever, V. N. Gladyshev, M. Vidal, P. Schow, J. Lehar, and F. P. Roth.** 2011. Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. Nat Methods **8:**159-164.

Publications since our last report for goal 2

1. **Algire, M. A., M. G. Montague, S. Vashee, C. Lartigue, and C. Merryman.** 2012. A Phase Variable Type III Restriction Modification system in *Mycoplasma mycoides* subsp. *capri*. BMC Microbiology. **In preparation**.
2. **Burgos, R., G. E. Wood, L. Young, J. I. Glass, and P. A. Totten.** 2012. RecA mediates MgpB and MgpC phase and antigenic variation in *Mycoplasma genitalium*, but plays a minor role in DNA repair. Infection & Immunity. **In press.**
3. **Dai, J., N. Assad-Garcia, N. Alperovich, R. Krishnakumar, R.-Y. Chuang, J. I. Glass, and S. Vashee.** 2012. Cre-Lox System to Recycle Markers in Mycoplasmas. Appl Environ Microbiol. **In preparation**.
4. **Grover, R. K., J. I. Glass, R. A. Kyle, D. F. Jelinek, X. Zhu, I. A. Wilson, D. Salomon, and R. A. Lerner.** 2012. Highly Selective Reactivity of Myeloma Immunoglobulins with Human Mycoplasma Antigens. **In preparation.**
5. **Juhas, M., L. Eberl, and J. I. Glass.** 2011. Essence of life: essential genes of minimal genomes. Trends Cell Biol **21:**562-568.
6. **Karr, J. R., J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival, N. Assad-Garcia, J. I. Glass, and M. W. Covert.** 2012. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell. **In press**
7. **Noskov, V. N., R. Y. Chuang, D. G. Gibson, S. H. Leem, V. Larionov, and N. Kouprina.** 2011. Isolation of circular yeast artificial chromosomes for synthetic biology and functional genomics studies. Nat Protoc **6:**89-96.
8. **Ramon, A., and H. O. Smith.** 2011. Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. Biotechnol Lett **33:**549-555.
9. **Wise, K. S., M. J. Calcutt, M. F. Foecking, R. Madupu, R. T. DeBoy, K. Röske, M. L. Hvinden, T. R. Martin, A. S. Durkin, J. I. Glass, and B. A. Methé.** 2012. The Complete Genome Sequences of *Mycoplasma leachii* Type strain PG50T and the Pathogenic *Mycoplasma mycoides* subsp. *mycoides* Small Colony Biotype Strain Gladysdale J. Bacteriol. **Submitted**.