143

Metagenomic and Metaproteomic Analysis of Viruses from an Acid Mine Drainage System

Kimberly Pause Tucker[1]* (kpause@marine.usf.edu), Christine L. Sun,[2] Brian D. Dill,[3] Paul Wilmes,[2] Robert Hettich,[3] Brian C. Thomas,[2] Vincent Denef,[2] Nathan VerBerkmoes,[3] Luis Comolli,[4] Birgit Luef,[4] **Jillian F. Banfield**,[2] and **Mya Breitbart**[1]

[1]University of South Florida, Saint Petersburg; [2]University of California, Berkeley; [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [4]Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals (Abstracts 143-145): The aim of this project is to develop integrated genomic and proteomic (proteogenomic) methods to study virus-host dynamics in bioreactor-grown and natural microbial biofilm communities. We will utilize simultaneous metagenomic analysis of CRISPRs (the adaptive microbial immune system) and viruses to examine the co-evolution and dynamics of viruses and their microbial hosts. Due to virus-host specificity, viruses can change the microbial community structure by enriching for resistant strains. Proteomics will be used to examine metabolic interplay via monitoring of protein production before, during, and after viral infection, while genomic analysis of CRISPRs will be used to correlate these findings with changes in community structure. Because viral predation can undermine microbially-based technologies such as microbial fuel cells, bioethanol production, and environmental remediation, methods to study microbe-viral interactions are widely relevant to DOE missions.

Metagenomics has proven to be a useful tool for examining viruses in a variety of natural systems, revealing novel and diverse environmental viral communities. In systems with low species richness, metagenomic sequencing can reveal patterns in microbial diversity and evolution with unprecedented resolution. Concomitant genomic sampling of virus populations and clustered regularly interspaced short palindromic repeat (CRISPR) loci from the low diversity acid mine drainage (AMD) biofilm system from Richmond Mine (Iron Mountain, CA, USA) offers unique insight into virus-host interactions (see Sun *et al.* poster). Five viral genomes (AMDV1-5) were previously reconstructed from metagenomic sequence data from the UBA site, and linked to their potential hosts through the CRISPR loci. Here we report analysis of an additional novel archaeal virus genome, AMDV6, which was reconstructed from a total microbial community metagenome from the C75 site. It appears that there was a bloom of viruses occurring at that time, since many of the total metagenomic sequences were viral in origin (see Sun *et al.* poster). Annotation of the AMDV6 genome revealed open reading frames (ORFs) with functions related to DNA-binding, replication, and modification. Other regions of interest include a phage integrase, as well as small molecule binding domains. The AMDV6 genome also exhibits similarity to some of the other AMD viruses. The largest AMDV6 ORF (1138 aa) shows approximately 30% amino acid identity to hypothetical membrane-associated proteins of AMDV4 and AMDV3. The second largest AMDV6 ORF (672 aa) shares 31% amino acid identity with the DNA polymerase elongation subunit (family B) of AMDV4. The spacers of CRISPR loci from the Archaea concurrently inhabiting the C75 site are being examined to identify potential hosts for the novel AMDV6 virus.

Purification of viral particles from the biofilm prior to metagenomic sequencing allows for the description of the complete viral community, including viruses present at low abundance that might not have been sampled in the total community metagenome, and viruses that are not targeted by CRISPRs. Using a combination of mechanical disruption, filtration, and density-dependent centrifugation, we have developed a protocol to purify viruses from the AMD biofilm for genomic and proteomic analyses. Metagenomic sequencing of viruses purified from the UBA-BS site demonstrated that a high proportion of the sequences had similarity to known AMD viral sequences, confirming the success of the virus purification method. Amongst the sequences with similarity to previously described AMD viruses, many were similar to AMDV4, a virus that infects E-plasma. The majority of the AMDV4 sequences were more than 98% identical on the nucleotide level to sequences in the database, which was generated three years prior. This level of sequence conservation indicates that some of the viral populations in the AMD system are not changing considerably over time despite high virus-host contact rates in the biofilm and an active CRISPR-mediated defense system. The high proportion of virus sequences with similarity to the five previously described AMD viruses suggests that there are only a few dominant viral types in the system, which is consistent with the low microbial species richness at this site. Although the majority of the viral sequences were similar to previously described AMD viruses, a number of sequences were identified that had no similarities to either the NCBI non-redundant database or to the AMD-specific metagenomic database, demonstrating the discovery of novel viral types from metagenomic sequencing of the purified viral fraction. Since these viruses were not previously identified from the whole community metagenomes, it is possible that they are not targeted by CRISPRs. Continued sequencing of the purified viral

fractions will allow us to estimate the percent of the AMD viral community that is targeted by CRISPRs. In addition, reconstruction of the complete genomes of these novel viruses will enable examination of differences in the rates and mechanisms of genome evolution between CRISPR-targeted viruses and those that are not affected by CRISPRs.

Metaproteomic analysis on AMD viral fractions purified from the UBA-BS site demonstrated a 12-fold enrichment of identified viral proteins compared to total community proteomes. Some non-viral proteins were also enriched, including flagellar proteins, which is likely a consequence of their co-purification during the virus particle selection protocol. The majority of identified viral proteins were attributed to AMDV1, a bacteriophage of *Leptospirillum* groups II and III. The amount of variability and rapid evolution of viruses makes identification of viral proteins challenging, prompting the need for *de novo* predictions (see VerBerkmoes *et al*. poster). Inclusion of six-frame translations for the AMDV1 genome lead to a ten-fold increase in protein identification for this virus, demonstrating the need for database expansion to include sequence variants (frameshifts and single nucleotide polymorphisms). Concurrent metagenomic and metaproteomic analyses of purified AMD viral communities will significantly increase protein identification rates. To assess the amount of sequence variability amongst individuals in the viral community and aid in protein identification, the sequence variation in AMDV1 over space and time was examined by PCR and sequencing of gene 1. The AMDV1 population exhibited extensive single nucleotide polymorphisms in this gene, which are likely generated in response to the CRISPR defense. Sequences from the UBA-BS site clustered separately from sequences from the B-drift site, indicating spatial differentiation in the AMDV1.

In conclusion, purification of viral communities from the AMD biofilm has expanded our understanding of the diversity and evolution of viruses in this extreme environment. Metagenomic and metaproteomic analyses of the purified viral communities overcome the difficulties of obtaining sufficient coverage of viral genes/proteins from total biofilm samples. The high levels of single nucleotide polymorphisms and rapid evolution of viruses in the AMD system present a unique challenge for protein identification, which will be a focus of future work (see VerBerkmoes *et al*. poster).

Funding provided by DOE Genomics:GTL Program grant number DE-FG02- 07ER64505

144

Community Proteogenomic Analysis of Virus-Host Interactions in a Natural System

Christine Sun[1]* (christine_sun@berkeley.edu), Brian Thomas,[1] Vincent Denef,[1] Kimberly Pause Tucker,[2] Paul Wilmes,[1] Rodolphe Barrangou,[4] Chongle Pan,[3] Michael P. Thelen,[5] Robert Hettich,[3] Nathan VerBerkmoes,[3] Mya Breitbart,[2] and **Jill Banfield**[1]

[1]University of California, Berkeley; [2]University of South Florida, Saint Petersburg; [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [4]Danisco USA Inc., Madison, Wis.; and [5]Lawrence Livermore National Laboratory, Livermore, Calif.

Project Goals: See goals for abstract 143.

Bacteriophages and archaeal viruses (collectively referred to here as viruses) play critical roles in microbial evolution as they can shape the composition and functionality of microbial communities via predation of their hosts and promotion of lateral gene transfer. Few studies have examined virus-host dynamics and viral diversity in natural populations due to difficulties in the recovery of viral population genomic data and the lack of cultivation-independent methods to link viruses to their hosts. However, the recent identification of CRISPRs (clustered regularly interspaced short palindromic repeats) as adaptive, rapidly evolving microbial immune systems provides a crucial cultivation-independent connection between a CRISPR-containing host and its viruses. CRISPR regions consist of repeats separated by short spacer sequences that derive from the genomes of viruses and other mobile elements. Since spacers are incorporated into the CRISPR in a unidirectional manner, the CRISPR region also serves as a historical timeline of virus exposure. We have simultaneously analyzed virus populations and CRISPR loci of the bacterial and archaeal populations they target to study virus-host interaction dynamics in relative low diversity acid mine drainage (AMD) microbial communities. To complement comprehensive examination of the dynamics of the CRISPR loci and viruses in natural populations, we initiated research on an *in vitro* system involving *Streptococcus thermophilus* and its lytic phage. The objective of the *in vitro* simple system research is to calculate the rates and determine the mechanisms of virus and host co-evolution.

We have shown that CRISPR loci in microbial genomes assembled from AMD community genomic datasets are extremely dynamic genomic regions that undergo rapid gain and loss of spacers. Via 454 FLX pyrosequencing of amplified CRISPR loci from two *Leptospirillum* Group II bacterial populations (UBA and 5way type), 419,351 total spacers were recovered. As 454 pyrosequencing generates reads with higher error rates relative to Sanger reads, spacer sequences were placed in groups that share at least 80% identity over 80% length, resulting in 2,841 and 649 spacer groups in the 5way and UBA type *Leptospirillum* Group II, respectively. The rarefaction curves for both samples using collapsed spacer groups demonstrate no approach to

saturation, implying a large diversity of CRISPR spacers in each population. Spacers were used to identify sequences from a viral population that targets *Leptospirillum* Group II (AMDV1). By mapping spacer sequences back onto assembled viral genomes, the "CTT" CRISPR PAM (proto-spacer adjacent motif involved in spacer selection) was identified. The presence of the motif is the only factor that appears to impact selection of the spacer sequence (i.e., there is no significant bias toward the coding vs. non-coding strand). We have identified many cases where multiple spacer sequence variants were sampled from the same locus in the diverse AMDV1 population. Spacers also target the *Leptospirillum* Group II genome, mostly in phage/plasmid-like regions. Interestingly, more 5way unique spacers target the UBA type *Leptospirillum* Group II than the 5way type. This may further support the previously suggested recent lateral transfer of the CRISPR/Cas system from the UBA to the 5way genome. Interestingly, spacers unique to 5way target transposons only located in the UBA *Leptospirillum* Group II type but not to the 5way type.

We compared the *Leptospirillum* Group II locus across seven time points from three locations and spanning five years. Notably, most spacers in the first fifth of loci are conserved in both the UBA and 5way *Leptospirillum* Group II genotypes across all samples, indicating that spacer loss occurs primarily in the middle of loci. In some cases, dissimilarities in spacer context indicate multiple independent sampling events at the same PAM. Five samples from one location did not show unidirectional locus expansion along the 2-year time series, suggesting that the predominant process is selection from among coexisting subpopulations.

At one time point, sequences recovered from one biofilm included a large representation of reads from virus populations. Dominant among these is a population (AMDV3b) related to virus previously described by metagenomics, AMDV3. Other populations present at lower levels include the bacteriophage AMDV1, viruses related to AMDV4, and new archaeal viral populations including AMDV6 and another variant of AMDV3 (AMDV3c). The AMDV3b population was very deeply sampled (>800 times coverage), providing a novel view of natural virus population structure. Levels of sequence heterogeneity and differences in gene content in the AMDV3b population are highly variable across the genome. Extensive heterogeneity is localized in a very large protein (~ 2,000 amino acids) of unknown function that we suspect is a tape measure protein involved in tail or capsid assembly, and thus host specificity. We detect extensive nucleotide polymorphism and small groups of nucleotide insertions/deletions in the gene for this protein, most of which leave the flanking amino acid sequence in frame. There is clear evidence for extensive, fine-scale homologous recombination (tens to hundreds of nucleotide blocks) amongst sequence variants and for construction of protein variants by mix and match involving large, distinct sequence modules.

On a shorter time scale using the *in vitro* system of *Streptococcus thermophilus* and its virus, CRISPR loci and viral sequences were examined at two time points separated by around 42 to 51 generations. No spacers in the original CRISPR loci were able to target the virus. However, at the second time point, new spacers were discovered in two of four CRISPR loci in *Streptococcus thermophilus*. At least 50 of these new spacers match the original viral population perfectly. In at least four instances, the virus population had mutated by single nucleotide polymorphisms within the proto-spacer or the PAM in order to escape the new spacers.

All evidence from both systems indicate that both CRISPRs and viruses are co-evolving rapidly, in an "arms-race" that requires continual acquisition of new spacers to target viral sequences that have been modified by mutation and sequence shuffling. An understanding of virus-host interaction dynamics is generally applicable and directly relevant to the DOE mission through the implications for the maintenance of stable biotechnologies, including bioremediation and industrial bioenergy production.

Funding provided by DOE Genomics:GTL Program grant number DE-FG02-07ER64505

145

Advanced Proteomics Methods to Identify Low Level Viral Signatures in Isolate and Environmental Samples

Nathan VerBerkmoes[1]* (verberkmoesn@ornl.gov), Brian Dill,[1] Jacque Young,[1] Kimberly Pause Tucker,[2] Christine Sun,[3] Brian Thomas,[3] Vincent Denef,[3] Paul Wilmes,[3] Patricia Carey,[1] Manesh Shah,[1] Chongle Pan,[1] Robert Hettich,[1] Rodolphe Barrangou,[4] Mya Breitbart,[2] and Jill Banfield[3]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]University of South Florida, Saint Petersburg; [3]University of California, Berkeley; [4]Danisco USA Inc., Madison, Wis.

Project Goals: See goals for abstract 143.

Genomic and proteomic methods have been developed to interrogate ecological interplay among bacterial and archaeal members of natural microbial communities. The relative simplicity of the acid mine drainage (AMD) biofilm system has made it a prime model for community genomic and proteomic approaches to investigate the ecology of a natural microbial community. While much progress has been made to begin to unravel the ecological and genetic interaction of bacterial and archaeal species in the AMD system, less has been discovered concerning the activity and ramifications of viruses in the AMD biofilms. Interestingly, most AMD bacteria and archaea have active CRISPRs, a microbial system which can convey immunity against viral attack. Viruses (bacteriophage and archaeal viruses) are ubiquitous in microbial communities, and have dramatic effects on species-level and community-level dynamics. By controlling fluctuation of microbial populations, initiating genetic exchange between populations, and directly altering

microbial physiology, viruses can affect carbon, nutrient, and metal cycling.

In situ genomic and proteomic analyses of viruses pose a daunting challenge due to the difficulty of obtaining sufficient sequence coverage given the low abundance of viruses relative to the bacterial and archaeal constituents and the high mutational rate exhibited by viruses. Our aim is to adapt community genomic and proteomic approaches to identify proteins from viruses in order to expand our understanding of physiological, ecological, and genetic viral-microbial interactions. To reach this aim we are developing mass spectrometry based proteomics techniques to confidently identify viral proteins in complex matrices. To develop these techniques, we are using two model systems: the *in vitro* biotechnology system of *Streptococcus thermophilus* and its lytic phage as well as the AMD biofilm community. The first system is well characterized genomically and provides an ideal model system to develop and test proteomic methodology because of the high concentrations of virus particle and host cells. However, although the minimal media is relatively straightforward, the milk proteins create a significant masking issue. The AMD biofilms are the best characterized microbial community via proteogenomic techniques to date, with the presence of viruses confirmed via genomic approaches and imaging). The challenge associated with detection of viral proteins in this system is their low natural abundance and rapidly evolving genomes and thus proteomes.

In the study of the model *Streptococcus thermophilus* and its lytic phage, our first focus is a time series of infection followed by viral enrichment and proteome analyses. Our goal is to fully determine the proteins present in the free-living virus versus proteins specifically expressed in the host during infection. We also hope to determine the host response to the viral onslaught. *S. thermophilus* TSH-466 grown in LM17 media was infected with D2972 phage (M.O.I=.01). At times 0, 30 minutes, 1 hour, 2 hours, 4 hours, and 24 hours post-infection, aliquots were taken from the culture and centrifuged (10,000 g) for the cellular fraction, then the supernatant enriched for virus via PEG precipitation. Cellular fractions were measured by 2d-LC-MS/MS and searched against the host and viral genomes. We have seen a clear enrichment of the intact virus in the later time points with deep coverage over all known structural proteins and the detection of many unknown viral proteins as well. For the whole cell fraction at early time points we identify ~50% of the host proteome with a major increase in viral proteins over time. We are currently analyzing the datasets to understand the host response to viral attack. With the model system we are also developing chemical and antibody depletion technologies to remove milk proteins effectively allowing us to monitor the microbe-viral interactions in milk, the biotechnology media of interest.

In the course of the AMD project we have characterized over 50 unique AMD community samples. We have built a new predicted proteome database which contained all of the predicted viral proteins to date as well as their variants, translated in all 6 open reading frames. We appended this database to the current AMD community predicted proteome database and queried our archived datasets. The new search identified significantly more viral proteins than the previous database. From this informatics exercise we are able to detect viral protein signatures across time and space with high accuracy using existing archived proteomics datasets. Currently we are focusing on the deep proteome characterization of a series of samples from a single location collected over a two year interval (C75 time series). One C75 sample captured a viral bloom (Banfield group). The detection of proteins from such viral blooms in metagenomically characterized samples should be possible. We are employing two new technology approaches to achieve deep detection of the viral signature. The first technology advance is the use of intact protein separations prior to digestion and 2d-LC-MS/MS. The second is the use of a new rapid scanning LTQ-Orbitrap Velos, which should provide a dramatic increase in proteome depth compared to LTQ-Orbitrap. One of the most promising methods to enhance detection of viruses is density gradient centrifugation of viruses extracted from AMD biofilms. These enriched viral protein samples have led to the highest detection rate of viral proteins from AMD biofilms to date. We are currently analyzing a new enrichment set from a time series. Finally we are using *de novo* proteomic sequencing methods to attempt to identify viral strain variants that would have otherwise escaped detection.

Funding provided by DOE Genomics:GTL Program grant number DE-FG02- 07ER64505. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

146

Integrating Experimental and Computational Approaches to Enhance Proteogenomic Characterizations of Natural Microbial Communities

Robert L. Hettich[1] (hettichrl@ornl.gov), Nathan VerBerkmoes,[1] Chongle Pan,[1] Brian Erickson,[1] Brian Dill,[1] Alison Russell,[1] Manesh Shah,[1] Doug Hyatt,[1] Ryan Mueller,[2] Vincent Denef,[2] Paul Wilmes,[2] Brian Thomas,[2] Brett Baker,[2] Korin Wheeler,[3] Steve Singer,[3] Michael Thelen,[3] and Jillian Banfield[2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]University of California, Berkeley; and [3]Lawrence Livermore National Laboratory, Livermore, Calif.

Project Goals: To establish a proteogenomic approach for elucidating how microbial consortia assemble and respond to their environmental pressures. To this end, we have focused on a model environmental microbial community found in acid mine drainage to develop and demonstrate a MS-based proteomic approach to investigate the molecular level activities of microbial consortia. Whole community genomics serves as the underpinning core for the measurements and evaluations of microbial consortia.

The characterization of natural microbial communities at a systems biology level can be addressed in significant detail with a technology platform based on high performance mass spectrometry interfaced closely with computational approaches for mining the raw MS data, compiling protein results, and finally integrating with physiological and genomic information. To this end, we have focused on a model environmental microbial community found in acid mine drainage to develop and establish a proteogenomic approach for elucidating how microbial consortia assemble and respond to their environmental pressures. Whole community genomics serves as the underpinning core for the subsequent measurements and evaluations of microbial consortia.

We have greatly expanded the range of temporal and spatially resolved AMD samples measured by proteogenomic techniques, and have analyzed over 50 distinct AMD biofilms to date, with an average of 2,000-4,000 protein identifications per site. The goal of this work is to explore the genomic and proteomic diversity of the AMD system. The resulting data has significantly expanded the molecular-level view of the AMD microbial system by revealing a deeper level of molecular information for not only the bacteria, but also archaea, nanoarchaea, and viruses as well. Recent work has focused on quantitative environmental proteomics to examine the effects of inter-species interactions on population physiology in natural microbial communities as ecological succession proceeds. For this work we sampled three defined developmental stages (early, intermediate, and late) in triplicate from the same natural environment and measured soluble and membrane proteome fractions for changes in the abundances of proteins from dominant and sub-dominant community members via label free and metabolic labeling quantitative proteomics. We identified approximately 6000 proteins across all samples and replicates (2522 from the membrane fraction and 3437 from the soluble fraction) from fifteen distinct genetic populations of this community. Biological replicates demonstrated high concordance, which deteriorated slightly as communities became more complex. The transition from early developmental stage biofilms to late developmental stage biofilms coincides with a physiological switch in the dominant population of these communities. Functional biases in early developmental stages pertain to rapid growth and protein synthesis (*i.e.*, cell division, ribosome assembly, transcription), whereas biases in the late developmental stage include functions for building block biosynthesis (AA, CHO, Lipids), chemical sensing, and translation. This switch may reflect physiological constraints imposed on this dominant member by the arrival of and competition with secondary colonizers in late stages of biofilm development.

One outcome of the proteogenomic work is the observation of a substantial sequence divergence in the genomes of closely-related environmental microbes, mostly in the form of single nucleotide polymorphisms, which translate into single amino acid variants at the protein level. Clearly, it is important to characterize these amino acid variants in order to understand their functional roles. We have developed a database-searching algorithm, termed *Polyscan*, to identify amino acid polymorphisms by shotgun proteomics. *Polyscan* systematically mutates *in-silico* one residue at a time for every protein to generate tryptic peptides with polymorphisms, and then correlates mutated peptides to the acquired tandem mass spectra. To efficiently search this enormously expanded sequence space, *Polyscan* uses a binary search to assign spectra to candidate peptides by parent masses and then scores peptides with a fast preliminary scoring function followed by a slower, but more accurate, primary scoring function. *Polyscan* was used to analyze a proteome of a natural microbial community in acid mine drainage. In the data analysis, the mutations of N to D, Q to E, and A to S were excluded, as these mutations have the same mass shift as common PTMs on the original amino acids. A total of 1102 polymorphisms were identified at a false discovery rate of 1.2% in an AMD sample measured by high resolution LTQ-Orbitrap-MS. The frequency of these polymorphisms correlates well with their BLOSUM62 propensity. This provides support for the use of *Polyscan* to identify amino acid polymorphisms in shotgun proteomics datasets.

We also have undertaken a characterization of the metalloproteome from the AMD system. Metals play essential roles in cellular metabolism, with both beneficial and toxic effects. Proteins can bind metals in catalytic centers, utilize metals as structural elements, or serve as chaperones to safely transport metals to their appropriate location within a cell. In some microbes, metals can even drive an organism's metabolism, providing the necessary energy for life. Of great interest are the biochemical interactions between metals and proteins that drive geomicrobiological processes. By growing in molar concentrations of iron and millimolar concentrations of copper, zinc and arsenic, AMD microbial biofilms are ideal candidates for investigations into metal-influenced biogeochemical interactions. Investigations into the relationships between the AMD microbial community and the available heavy metals provide insights into the metabolic processes that enable life in this extreme environment. We have completed an initial assessment of the metal-binding proteome for an AMD microbial community. We employed selective enrichment of metal-binding proteins with immobilized metal affinity chromatography (IMAC) across seven different biologically active metals (copper, cobalt, manganese, magnesium, nickel, zinc, and iron) with a specific focus on identifying bound vs. unbound proteins in each case. On average, about 270 proteins were identified in each column fraction, with an average of 9% variability between replicates. In every column except iron, there were more proteins identified in the unbound fraction, with about 54% of bound proteins having no known function. The largest groups of specific metal binding proteins are involved in translation, ribosomal structure, and biogenesis, or post-translational modification, protein turnover and chaperones.

In total, our proteogenomic methodology provides a powerful approach for the systems biology interrogation of natural microbial communities, and should be broadly applicable for microbial systems of relevance for environmental remediation and bioenergy production.

This research sponsored by the U.S. DOE-BER, Genomics Program. Oak Ridge National Laboratory is managed by

UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

147

Understanding Carbon Cycling in a Model Microbial Community by Integration of Stable Isotope Probing, Metabolomic, and Transcriptomic Data With Proteogenomic Analyses

Curt R. Fischer[1]* (curt_f@berkeley.edu), Brett J. Baker,[1] Paul Wilmes,[1] Chris Belnap,[1] Chongle Pan,[2] Shufen Ma,[1] Trent Northen,[3] Nathan C. VerBerkmoes,[2] Benjamin Bowen,[3] Susan Spaulding,[1] Pepper Yelton,[1] Nicholas Justice,[1] Ryan S. Mueller,[1] Vincent J. Denef,[1] Brian C. Thomas,[1] Steven W. Singer,[3,4] Michael P. Thelen,[4] Robert L. Hettich,[2] and **Jillian F. Banfield**[1]

[1]University of California, Berkeley; [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [4]Lawrence Livermore National Laboratory, Livermore, Calif.

Project Goals: Our recent aims have included development of methods to achieve a comprehensive, molecularly-resolved, multi-scale understanding of carbon and energy flow in a model chemoautotrophic microbial community of tractable complexity, in order to lay a foundation for work in more complex environments. The first level of our analysis is to define the metabolic potential of coexisting natural populations through metagenomic sequencing. The second level of our analysis includes whole-community shotgun proteomics and mass-spectrometry-based metabolomics to identify key chemical and biochemical intermediates of carbon and energy transfer in these communities. The third level of our analysis aims at systems-level studies of the community to track the flow of carbon and energy over a range of environmental conditions and relies on cultivation of the entire community.

Microbial communities drive global biogeochemical cycling of energy and carbon, and a molecular understanding of how they respond to environmental stresses can inform predictions of how cycles will change in the future. However, many natural ecosystems are too complex to currently study in molecular detail. Our aim is to develop methods to achieve a comprehensive, molecularly-resolved, multi-scale understanding of carbon and energy flow in a model chemoautotrophic microbial community of tractable complexity, in order to lay a foundation for work in more complex environments. Our studies toward this end can be broken into three levels. The first level of our analysis is to define the metabolic potential of community members through metagenomic sequencing. In past GTL-sponsored research, we have assembled high-quality strain-resolved genomes for the three bacterial iron-oxidizing primary producers of our model community and draft genome assemblies for six archaeal mixotrophs of the order *Thermoplasmatales*. Recent functional annotations indicated that all of these archaea are facultative anaerobes and have complete pathways for glycolysis, beta-oxidation, and polypeptide degradation. We also assembled near-complete genomes for three deeply-branched, low abundance and uncultivated lineages of Euryarchaea (referred to as ARMAN-2, 4 and 5). The small genome size, approximately 1 Mb, is consistent with their small cell size (~500 nm in diameter). A shorter than average gene length allows for a 10% higher coding density than is achieved by other genomes of comparable size. Two of the ARMAN groups have complete glycolytic pathways and the third has a beta-oxidation pathway for fatty acid utilization, thus all are likely capable of heterotrophic growth.

The second level of our analysis includes whole-community shotgun proteomics and mass-spectrometry-based metabolomics to identify key chemical and biochemical intermediates of carbon and energy transfer in these communities. Due to the complexity of the metabolome and lack of appropriate reference spectra, it was necessary to analyze metabolomic data without feature identification. Over 8,000 metabolic features were identified from a collection of 15 samples, representing 7 environment types. Proteomic data, in which detected peptides are matched to proteins predicted from assembled genomes, were also collected for each sample, and metabolomic features were co-analyzed with proteomic data. Correlation and clustering analyses revealed two large, well-resolved groups of metabolites. Each group strongly correlates with proteins from only one of the two dominant bacteria. This finding suggests that evolutionary divergence of these two co-existing bacteria has resulted in divergent metabolite profiles, possibly reflecting differences in genome regulation, kinetics of orthologous enzymes, and the presence of organism-specific metabolites.

The third level of our analysis aims at systems-level studies of the community to track the flow of carbon and energy over a range of environmental conditions and relies on cultivation of the entire community. Laboratory growth of biofilms allows both precise control over environmental conditions and the use of stable isotope as labels or tracers. Our culture system uses continuous-flow peristaltic pumps to move an acidic, iron-rich growth medium through custom-designed open-air Teflon channels situated in a temperature-controlled chamber at 40 °C. We have observed gross metabolic rates of between 1 and 15 W/m^2 in the bioreactor, as calculated from Fe^{2+} oxidation rates. This is comparable to, and sometimes higher than, field-observed values of 0.6 to 3.7 W/m^2 for our model system, and additionally is similar to rates of phototrophic primary production in terrestrial ecosystems.

We have validated that we can track and quantify stable isotopic tracers in the proteomes and metabolomes of our model community. In one proof-of-principle experiment, community proteomes were extracted from two bioreactors, one of which was cultivated with a 50:50 mixture of $^{15}NH_4^+$ and $^{14}NH_4^+$ as the sole nitrogen source, and the other with $^{15}NH_4^+$ as the sole N source. We used the Polyscan algorithm to identify both the sequence and ^{15}N isotopic

enrichment ratio of 1449 peptides and 2404 peptides in the two samples, respectively. The median peptide isotopic enrichment in the isotopically pure $^{15}NH_4^+$ sample was 99%, with the 20th and 80th percentiles being 98% and 100%, illustrating that the vast majority of calculated peptide isotopic enrichments were similar to the experimentally imposed value. The median peptide isotopic enrichment in the 50% labeled sample was 49%, with the 20th and 80th percentiles being 47% and 49%, showing that even the complex mixtures of peptide isotopomers that arise in partially labeled samples can be correctly identified. Separately, in a metabolomics proof of principle experiment, we quantified deuterium incorporation into phosphatidylethanolamine lipids from consortia grown in medium with 10% deuterated water using mass-spectrometry-based metabolomics. The experiments indicate that stable isotopic tracers can be quantified using traditional proteomic and metabolomic mass spectrometry techniques, and that these techniques can be used to trace the flow of stable isotopic tracers throughout community metabolomes and proteomes.

Taken together, the data are beginning to provide a highly integrated picture of carbon cycling in our model community. Carbon is initially fixed by two dominant bacteria (*Leptospirillum* spp. of the Nitrospirae), which also produce cellulose-based polysaccharides and osmolytes that become substrates for the growth of archaea, low abundance bacteria, and fungi. Archaea grow heterotrophically or mixotrophically by beta-oxidation of fatty acids or via other complex organic carbon degradation pathways. Fungi appear in later developmental stage aerobic biofilms. We speculate that fungi help catalyze the biodegradation of cellulose and other polysaccharides in the sulfuric acid-rich pH 1 environment of our model ecosystem, and so we are also targeting these organisms for metagenomic and metatranscriptomic sequencing. Less dominant community members, including actinobacteria and the ultra-small ARMAN archaea are ubiquitous and have intimate interactions that likely involve C transfer with as yet unidentified cell wall-less archaea. Sulfite- and possibly Fe^{3+}-reducing archaea and bacteria may play roles in anaerobic metabolism of organic carbon. Initial evidence in this direction comes from microelectrode-based analysis of waters and sediments from mine and lab bioreactor samples, as well as initial proteogenomic characterization of biofilms sampled from submerged, anaerobic environments. Microelectrode measurements have revealed thiosulfate, sulfite and polysulfides which we believe may be involved in community energy cycling. We have detected shifts in community population structure in sunken biofilms, with increased levels of *Sulfobacillus* and *Thermoplasmatales* archaea relative to surface biofilms. Future transcriptomic, proteomic, metabolomic and stable-isotope labeling experiments will test these hypotheses. Both the methods and the novel integrative bioinformatic architecture that we have developed may become extensible to other, more complex ecosystems as the metagenomics knowledge base for these ecosystems is developed.

Funding provided by DOE Genomics:GTL Program grant number DE-FG02-05ER64134.

148

Sub-Proteomic Analyses and Structural Modeling of Proteins Expressed in Acidophilic Microbial Communities

Yongqin Jiao[1]* (jiao1@llnl.gov), Pepper Yelton,[2] Patrik D'haeseleer,[1] Adam Zemla,[1] Brian Dill,[3] Nathan Verberkmoes,[3] Robert Hettich,[3] **Jill Banfield**,[2] and **Michael Thelen**[1] (mthelen@llnl.gov)

[1]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.; [2]Dept. of Environmental Science, Policy, and Management, University of California, Berkeley; and [3]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: To gain a comprehensive understanding of an extremophile microbial community, its member interactions, biofilm formation, and geochemical impact using an approach that integrates metagenomics, MS proteomics, and biochemistry.

EPS matrix proteome

Extracellular polymeric substances (EPS) are major structural components in biofilms, providing spatial organization and structural stability to the microbial community. The exact composition of EPS varies substantially between different biofilms, and it remains a considerable challenge to provide a complete biochemical profile for interpreting how changes in EPS constituents affect community organization and development. In this study, we performed proteomic analysis to identify proteins that are present in the EPS matrix, and discuss correlations between EPS proteins and their potential microbial functions.

As part of our ongoing investigation of acidophilic microbial communities that generate acid mine drainage (AMD) at the Richmond Mine in Northern California, we examined proteins extracted from a mid-developmental stage biofilm (DS1) and a mature biofilm (DS2). Each biofilm was separated into EPS, whole cell, and mine solution fractions, and analyzed by MS proteomics following trypsin digestion. To determine proteins enriched or associated with the biofilm matrix, we compared the relative abundance of proteins identified in EPS with whole cell and mine solution proteins as a function of biofilm developmental stage.

In order to augment the existing functional annotation of the proteins, we applied a variety of standard subcellular localization prediction tools, and assigned putative enzymatic functions (EC numbers) using PRIAM. BLAST scores from matches within the Carbohydrate Active enZyme database (CAZy) were also used to assign protein families and additional EC numbers. The composition of the proteome in the EPS and mine solution fraction is very different from that in the cellular fraction, with more than 80% of the proteins detected in the cellular fraction underrepresented or undetectable in the EPS or mine solution. In contrast, predicted periplasmic and extracellular proteins are

over-represented by 3-7 fold in the EPS and mine solution compared to the whole cell fraction. Likewise, predicted outer membrane proteins are ~3-fold over-represented in the EPS. Some of these over-represented categories are dominated by a few abundant proteins. For example, much of the overabundance of extracellular proteins in the EPS fraction can be attributed to highly abundant putative flagellin proteins.

Enzymes overall are under-represented in the EPS and mine solution proteome: whereas around 50% of the peptide count for the cell fraction is associated with proteins that have an enzyme annotation, only around 10% of the peptide count of the mine solution and EPS fraction appears to be enzymatic. We also noticed over/under-representation of specific enzymatic functions (EC numbers) in different fractions. For example, among the most abundant enzyme functions, a protein disulfide-isomerase (EC 5.3.4.1), a peptidase (EC 3.4.24.64), and an adenylate kinase (EC 2.7.4.3, 2.7.4.10) are over-represented in the EPS and mine solution compared to the whole cell fraction, although most enzymes (e.g. carbohydrate-active) are estimated at much lower abundance.

Differences between DS1 and DS2 for each fraction are smaller than the differences between the fractions, and are likely due to variation in community composition, in addition to differential protein secretion and degradation during biofilm maturation. Some disparity in overall numbers of proteins belonging to the families of Glycoside Hydrolases, Glycosyl Transferases, and Carbohydrate Esterases were observed between the DS1 and DS2 proteomes, and may shed light on the maturation and carbohydrate composition of the EPS matrix.

The biofilm archaea

To analyze proteins expressed specifically from the Archaeal members of the biofilm community, we used proteomics, gene synteny, and structural homology. Gene order conservation (synteny) over evolutionary time is relatively rare. Genomes tend to be shuffled rapidly except in close relatives and in circumstances where neighboring genes are functionally related. This leads to a strong correlation between synteny and all measures of evolutionary distance. Preservation of synteny over large evolutionary distances should be weighted more strongly in gene function prediction because it is likely to not be due to chance but to result from negative selection against rearrangements. Based on this principle we developed our own weighted synteny-based approach to more accurately annotate poorly annotated genes in a group of co-occurring AMD Archaea. These organisms comprise one lineage consisting of four newly assembled genomes (A-, E-, G-, and I-plasma) in addition to five previously sequenced organisms of the order *Thermoplasmatales*.

Because tertiary structure is of primary importance in protein function, structural homology modeling was used to improve upon synteny-based annotations. Using a modified, high throughput version of an automated homology modeling system AS2TS, we predicted full or partial structures of a number of proteins belonging to these Archaea. Combined with synteny data, models of these poorly annotated proteins have allowed us to better predict protein function and thus further understand the role of these Archaea in the AMD community. Our results indicate that two of these organisms, A- and I-plasma, are capable of making the molybdopterin guanine dinucleotide cofactor and using it in anaerobic energy conservation metabolism. Proteomics confirms that some of these proteins are expressed in anaerobic sunken biofilms. This may help these species to thrive in adverse anaerobic conditions where other AMD Archaea cannot.

We also used this approach to annotate a region of proteins of unknown function found in only one of the AMD Archaea, G-plasma. This region of nine neighboring genes is notable in that most of the genes have no homologs in any of the other Archaea in this lineage or in available gene or protein databases. Nevertheless, proteomic data suggest that eight of the nine genes are expressed in certain conditions. In fact, one of the most highly detected G-plasma proteins is among this group. Protein modeling has been utilized to assign function to some of these unusual proteins. As a whole, these proteins appear to be involved in some type of protein or peptide modification and export or import. They model to reference proteins including a sulfurtransferase, a peptidase, a glutamyl-transpeptidase, and a large pore ABC transporter-like protein. Based on these annotations and proteomic data, we infer that this gene cluster plays an important role in G-plasma's differentiation from its co-occurring close relatives. The novel genes may be exporting peptides for a number of purposes to that effect, including for quorum sensing or as peptide antibiotics that make G-plasma more competitive.

This work was funded by the DOE Genomics:GTL Program grant number DE-FG02-05ER64134, and performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory and University of California, Berkeley under Contract DE-AC52-07NA27344. LLNL-ABS-420907.

149

Bottom Up Genome-Scale (BUGS) Modeling of Microbial Communities Influencing the Fate and Transport of Groundwater Contaminants

R. Mahadevan[2]* (krishna.mahadevan@utoronto.ca), Kai Zhuang,[2] Nadeera Jayasinghe,[2] Srinath Garg,[2] Jiao Zhao,[2] Tae Hoon Yang,[1] Jun Sun,[1] Yilin Fang,[3] Tim Scheibe,[3] Melissa Barlett,[4] Maddalena V. Coppi,[4] and **Derek Lovley**[4] (dlovey@microbio.umass.edu)

[1]Genomatica Inc., San Diego, Calif.; [2]University of Toronto, Toronto, Ontario; [3]Pacific Northwest National Laboratory, Richland, Wash.; and [4]University of Massachusetts, Amherst

Project Goals (Abstracts 149-157): The primary goals of this project are to develop genome-based experimental approaches to elucidate the function of anaerobic microbial communities involved in groundwater bioremediation and novel bioenergy processes and to develop genome-scale models of these communities in order to better understand and optimize bioremediation and bioenergy applications. One area of focus is the subsurface microbial community involved in the in situ bioremediation of uranium-contaminated groundwater. The other focus area are the microorganisms that can exchange electrons with electrodes, which can be employed in the bioremediation of subsurface environments, harvesting electricity from organic matter, and fixation of carbon dioxide into desirable organic products. These studies are demonstrating that with the appropriate genome-scale models it is possible to predict the in situ growth and activity of the relevant microorganisms in the environments of interest solely from relevant geochemical data and that the in situ metabolic state of the microorganisms can be diagnosed via analysis of gene expression of the subsurface community. It is expected that these studies will provide computational tools that can be used to predict the response of the microbial community to environmental manipulations, or manipulation of the genome of the relevant organisms, allowing rational optimization bioremediation and bioenergy applications via environmental or genetic engineering.

Studies at the Department of Energy field study site in Rifle, Colorado have demonstrated that a diversity of microorganisms are competing for resources in the subsurface and that the outcome of this competition can greatly impact the fate and transport of groundwater contaminants such as uranium. In order to better understand these interactions of microorganisms and their environment and their influence on uranium mobility, we have initiated a process known as BUGS (Bottom Up Genome-Scale) Modeling which has the ultimate goal of developing the ability to predict the efficacy of various bioremediation approaches, whether engineered or natural attenuation, prior to their implementation in the field.

In the BUGS Modeling approach, the most abundant, metabolically active microbes in the subsurface are identified with molecular techniques. Representatives of these key members of the subsurface microbial community are recovered in culture and genome scale metabolic models are developed for each microorganism. The genome-scale metabolic models are then coupled with geochemical and hydrological models to determine whether the models can predict the geochemical consequences of the response of the microbial community to various environmental perturbations in laboratory and field studies.

In previous studies it was demonstrated that the BUGS Modeling approach successfully predict the changes in uranium concentrations in groundwater at the Rifle site following the addition of acetate to the groundwater. This was a relatively simple test case because over 90% of the active microbial community was comprised of *Geobacter* species over the time frame modeled. It was also demonstrated that BUGS Modeling could describe the competition for resources between *Rhodoferax* species, another important group of Fe (III) reducers at the site, and *Geobacter* species and predict the relative distribution of these organisms under different geochemical regimes during natural attenuation as well as engineered bioremediation.

The initial bloom of *Geobacter* and successful removal of uranium from the groundwater following acetate addition at the Rifle site is followed by an increase in sulfate-reducing microorganisms that are ineffective in uranium reduction. In order to better and understand this phenomenon and identify strategies for managing this aspect of the bioremediation the interaction between *Geobacter* and sulfate reducers was modeled. Simulations demonstrated that no matter what the starting conditions, *Geobacter* species were able to dominate the initial phase of engineered bioremediation with very little change in timing of the onset of sulfate-reduction. The field-scale simulation was able to accurately predict the data from field experiments. The results suggest that the observed succession of *Geobacter* and sulfate reducers can primarily be attributed to differences in growth rates rather than other factors, such as competition for acetate. The simulations suggest that addition of Fe (III) to the subsurface can extend the time frame for effective uranium removal, but it is essential that the Fe (III) be added early in the bioremediation before sulfate reducers have had a chance to proliferate.

Another environment that strongly selects for *Geobacter* species is the surface of electrodes harvesting electricity from organic matter. Studies in many laboratories have demonstrated that in *Geobacter* species emerge from innocula from a diversity of environments as the predominant microbes on the anodes of the most highly effective microbial fuel cells. Current production in microbial fuel cells is controlled by a variety of factors including rates of substrate consumption and the accumulation of protons within the anode biofilm. It is important to be able to model the spatial gradients of substrates and inhibitors and to account for these heterogeneities on metabolism throughout the biofilm and their impact on current production. *Geobacter sulfurreducens*

produces the highest current densities of any known culture. Therefore, the genome-scale metabolic model of *Geobacter sulfurreducens* was coupled with a biofilm model in order to further investigate metabolism in anode biofilms. With this coupled model it has been possible to determine what regions of the biofilm are limited for substrate under different anode potentials and to evaluate the effect of different maintenance energy requirements on maximum current production and biofilm thickness. A strain of *G. sulfurreducens*, designated KN400, was recently selected in adaptive evolution studies that has substantially higher current-production capabilities than the starting strain, DL1. When estimated differences in cell yield and maximum substrate uptake rates were included in the model, the model accurately predicted the thinner, yet more effective current-producing biofilms of KN400. The model also successfully predicted the impact on current production of genetically engineering a futile cycle in DL1 to increase respiration rates. Additional development of the model to account for the impact of proton accumulation within the biofilm is underway.

The importance of the *G. sulfurreducens* genome-scale metabolic model for predictive analysis and optimization of *in situ* uranium bioremediation and microbial fuel cells justifies further refinement of this model. In order to further validate the model and to gain insights into intracellular metabolic response of *G. sulfurreducens* towards environmental perturbations its metabolism was further analyzed with ^{13}C-based metabolic flux analysis to characterize growth under conditions in which the electron donor (either acetate or hydrogen) or the electron acceptor (either Fe (III) or fumarate) limited growth. The donor and acceptor variations gave rise to differences in the pathways for gluconeogenesis, tricarboxylic acid cycle activities, and amino acid synthesis pathways. For example, cells were able to utilize fumarate as both an electron acceptor and as an additional carbon source when it was provided with acetate. As a result, cells slightly elevated the metabolic fluxes in the tricarboxylic acid cycle, and gluconeogenesis was initiated by phosphoenolpyruvate decarboxylase. In contrast, direct conversion of acetyl-CoA to pyruvate was the main source for the gluconeogenesis when Fe (III) was provided as the electron acceptor. Furthermore, the net flux direction between acetyl-CoA and pyruvate was reversed with fumarate as electron donor compared to Fe (III) as electron acceptor. Fluxes in the tricarboxylic acid cycle were lower with hydrogen as an electron donor compared to when acetate was the electron donor. These results improve the ability of the model to represent the physiology of *Geobacter* species under a diversity of growth conditions.

Additional refinements to the *G. sulfurreducens* model included the development of a thermodynamics based metabolic flux analysis model that successfully identified reactions that are subject to regulatory control, consistent with gene expression data. A novel model of uranium reduction based on a recently discovered electron storage pathways was developed, which improved predictions of uranium reduction in column studies.

These studies demonstrate the value of the BUGS modeling approach for predictively modeling the response of microbial communities to geochemical gradients and environmental perturbations. It is expected that BUGS Modeling will have wide spread application for a range of environmental applications, including the response of microbial communities to climate change.

150

Feeding Microbes Electricity: Gene Expression and Deletion Analysis of Mechanisms for Electron Transfer from Electrodes to Microbes

Sarah M. Strycharz, Kelly P. Nevin* (knevin@microbio.umass.edu), Richard H. Glaven, and **Derek R. Lovley**

Dept. of Microbiology, University of Massachusetts, Amherst

http://www.geobacter.org

Project Goals: See goals for abstract 149.

A rapidly emerging area in bioenergy and bioremediation is the possibility of driving beneficial microbial processes with electrons supplied with electrodes. This is a reverse in electron flow from the more commonly considered microbe-electrode interaction of current production in microbial fuel cells. Potential applications for feeding microbes electrons include bioremediation of waters contaminated with organics, radionuclides, metals, and/or nutrients; as well as the production of fuels and chemicals. Electrons can be supplied from a diversity of sources including wind and solar energy. In order to optimize these applications it is important to understand how microorganisms can use electrons supplied with electrodes as an energy source.

We have identified the first pure cultures capable of directly accepting electrons from electrode surfaces. For example, *Geobacter* species have been identified that can catalyze the reduction of organic acids, nitrate, chlorinated solvents, and metals with electrons supplied from an electrode as the sole electron donor. Other electron acceptors and other microorganisms with practical applications will be revealed at the time of the meeting.

Initial mechanistic studies were carried out with *Geobacter sulfurreducens*. Electrons were supplied with a graphite electrode poised at -500 mV versus Ag/AgCl. Fumarate was provided as the sole electron acceptor. Gene expression in current-consuming cells growing as biofilms on the graphite surface was compared with gene expression in two alternative growth modes: 1) biofilms growing on the same graphite material, but with acetate as the electron donor, rather than electrons from the electrode; and 2) biofilms growing on the same graphite material, but with acetate serving as the electron donor and the graphite serving as an electron-accepting electrode. These alternatives were termed no-current cells and current-producing cells, respectively.

Surprisingly, microarray analysis revealed that gene expression patterns in the current-consuming cells was significantly different than the previously reported gene expression patterns in current-producing cells. For example, current-producing cells have high expression of *pilA*, the gene encoding the structural protein for the pili termed microbial nanowires that are associated with high conductivity through current-producing biofilms. Current-producing cells also have enhanced expression of outer-surface *c*-type cytochrome genes, most notably OmcZ, which is required for optimal current production. However, *pilA*, *omcZ*, and other genes that are highly expressed in current-producing cells were expressed at low levels in current-consuming cells.

Furthermore, deletion of genes for *pilA* as well as *omcZ* and several other outer-surface *c*-type cytochromes had no impact on the current-consuming capabilities of *G. sulfurreducens*. This contrasts with the observed inhibition in current production when *pilA* or *omcZ* are deleted.

The concept that electrons are directly transferred from electrodes to *G. sulfurreducens* is based on several lines of evidence that suggest that hydrogen gas produced at the electrode surface is not an intermediate for electron transfer from the electrode to the cells. If hydrogen was an important electron transfer intermediate then it would be expected that genes for subunits of the uptake hydrogenase, Hyb, would be more highly expressed in current-consuming biofilms than current-producing or no-current biofilms. The gene for HybS, which encodes the small Hyb subunit, did have slightly higher transcript abundance in current-consuming cells versus the no-current control. However, this increase was just above the threshold for significance and the expression of the other Hyb subunits essential for a functional uptake hydrogenase were not upregulated in the current-consuming cells. Furthermore, transcript abundance for hydrogenases in current-consuming cells was not higher than in current-producing cells. Thus, gene expression analysis suggests that hydrogen was not an important intermediate in electron transfer from electrodes.

In both comparative studies cells accepting electrons from an electrode had a greater transcript abundance for a gene (GSU3274) encoding a putative monoheme *c*-type cytochrome. Deletion of this gene completely inhibited electron transfer from electrodes, but had no impact on the capacity for electron transfer to electrodes.

The amino acid sequence of GSU3274 shares homology most closely with putative cytochrome *c* family proteins from *Pelobacter propionicus*, *Thioalkalivibrio* sp., *Leptothrix cholodnii*, *Rhodoferax ferrireducens*, and *Polaromonas sp.*, all of which appear to contain a signal peptide cleavage domain suggesting they are translocated to the outer membrane. However, GSU3274 does not have a signal peptide cleavage site and is predicted to be localized in the periplasm. Therefore, it is unlikely that this protein serves as an electrical contact between the cells and the electrode. The structure of the protein encoded by GSU3274 is predicted to be similar to a cytochrome c_2 from the photosynthetic bacterium, *Rhodopila globiformis*, with a high redox potential. Thus, a potential role for the GSU3274 cytochrome is to serve as an intermediary in electron transfer between the outer cell surface and the inner membrane.

These results suggest that *G. sulfurreducens* has mechanisms for transferring electrons from electrodes that are substantially different than those for transferring electrons to electrodes. Differential gene expression in current-consuming electrodes versus current-producing electrodes may be dictated by the simple fact that the applied electrode potential influences the range of proteins for which electrode-cell interaction is energetically favorable. Furthermore, once electrons are transferred across the inner membrane, the remaining steps in electron transfer to electrodes do not require mechanisms for energy conservation, merely a pathway for electrons to flow down a potential gradient. In contrast, the pathway for electron transfer from electrodes into the cell must be specifically linked to a mechanism for generating a proton-motive force. These are significantly different metabolic demands.

Comparative analysis of gene expression in other organisms capable of accepting electrons from electrodes is underway and are expected to yield further insights into this important pathway for microbe-electrode interactions.

151

Immunocytochemistry Reveals Novel Models for the Role of Outer-Surface Cytochromes in Electron Transfer to Electrodes and Metals

Ching Leang* (leang@microbio.umass.edu), Kengo Inoue, Xinlei Qian, Tünde Mester, Trevor L. Woodard, Kelly P. Nevin, and **Derek R. Lovley** (dlovley@microbio.umass.edu)

University of Massachusetts, Amherst

Project Goals: See goals for abstract 149.

There are multiple competing/complementary models for extracellular electron transfer in Fe (III)- and electrode-reducing microorganisms. Which mechanisms prevail in different microorganisms or environmental conditions may greatly influence which microorganisms compete most successfully in sedimentary environments or on the surface of electrodes and can impact practical decisions on the best strategies to promote Fe (III) reduction for bioremediation applications or to enhance the power output of microbial fuel cells.

The three most commonly considered mechanisms for electron transfer to extracellular electron acceptors are: 1) direct contact between redox-active proteins on the outer surface of the cells and the electron acceptor; 2) electron transfer via soluble electron shuttling molecules; and 3) the conduction of electrons along pili or other filamentous structures. Evidence for mechanism #1 includes the findings that Fe (III)- and electrode-reducing microorganisms display redox-active proteins on their outer surface and that deletion of the genes

for these proteins often inhibits, at least partially, Fe (III) reduction and electron transfer to electrodes. Evidence for mechanism #2 includes the ability of some microorganisms to reduce Fe (III) which they cannot directly contact associated with the accumulation of redox-active soluble substances as well as specific electron shuttling electrochemical signatures of some microorganisms growing on an electrode surface. Evidence for mechanism #3 is more circumstantial and many authorities have questioned the potential for electron transfer along filaments.

OmcS is one of the most abundant cytochromes that can readily be sheared from the outer cell surface of *Geobacter sulfurreducens* and gene deletion studies have demonstrated that it is essential for the reduction of Fe (III) oxide as well as electron transfer to electrodes under some conditions. Therefore, the localization of this important protein was further investigated with immunogold labeling and electron microscopy.

With fumarate as the electron acceptor OmcS was in low abundance and primarily localized on the outer surface of the cells during early to mid-log phase growth. However, in cultures from late-log or stationary phase, the gold particles appeared as strands emanating from the cells. At higher magnification it was apparent that OmcS was associated with filaments with the same diameter and length of the previously described microbial nanowires of *G. sulfurreducens*. *G. sulfurreducens* produces more OmcS when grown under electron acceptor-limiting conditions and the greater abundance of OmcS associated with filaments was readily apparent in these cells. Filaments adorned with OmcS were also abundant in cells grown with Fe (III) oxide as the electron acceptor. When a strain of *G. sulfurreducens* in which the gene for OmcS was deleted was examined in the same manner there were no gold particles associated with the filaments.

This is the first description of cytochromes specifically associating with the filaments of a microorganism capable of extracellular electron transfer. The concept of cytochromes aligning with filaments of *Shewanella oneidensis* was previously proposed from indirect evidence and subsequent studies demonstrated that the cytochromes that were proposed to be aligned with filaments were randomly distributed within the extracellular matrix or associated with the outer cell surface rather than associated with filaments.

The dense packing of OmcS along the filaments suggests that cytochrome-to-cytochrome electron transfer might greatly facilitate electron transfer along the filaments.

These results do not negate the possibility that electrons may be transferred along the length of *G. sulfurreducens* pili without the need for cytochromes, but offer an alternative, possibly complimentary mechanism for long-range electron transfer.

Unlike OmcS, the outer-surface, *c*-type cytochrome OmcZ is not required for Fe (III) oxide reduction, but it is essential for optimal current production by *G. sulfurreducens*. In fumarate-grown cells OmcZ was dispersed throughout the extracellular matrix surrounding the cells of the biofilms that accumulated at the bottom of the culture tubes. When *G. sulfurreducens* grew as a biofilm on a graphite electrode that served as an anode and the sole electron acceptor for growth, OmcZ was highly concentrated at the biofilm-electrode interface. Controls in which the biofilm was grown on the same graphite material, but with fumarate as the electron acceptor, did not have accumulations of OmcZ at the electrode, corresponding with the inability of fumarate-grown biofilms to produce current. Although OmcS was also detected at the biofilm-electrode interface, it was also distributed throughout the biofilm. The specific localization of OmcZ at the anode surface under current-producing conditions, coupled with the previously published finding that deleting the gene for OmcZ dramatically increases the resistance of electron exchange between the anode and the biofilm, suggests that OmcZ may serve as an electrochemical gate promoting electron transfer from *G. sulfurreducens* biofilms to the anode surface. This proposed mechanism is consistent with the finding that the multiple hemes in the recently purified OmcZ protein have a wide range of mid-point potentials that may facilitate electron transfer at the divergent anode potentials that have been noted under different conditions.

The discovery of apparent novel roles for electron transfer for outer-surface *c*-type cytochromes in *G. sulfurreducens* emphasizes the relatively primitive understanding of extracellular electron transfer. Further investigations are warranted if the full potential of these organisms in practical applications such as bioremediation and conversion of organic compounds to electricity are to be realized.

152

Bioinformatic Analysis of Gene Regulation in the Metal-Reducing Bacterial Family *Geobacteraceae*

Julia Krushkal[1]* (jkrushka@utmem.edu), Ronald M. Adkins,[1] Yanhua Qu,[1] Jeanette Peeples,[1] Sreedhar Sontineni,[1] Ching Leang,[2] Peter Brown,[2] Nelson D. Young,[2] Toshiyuki Ueki,[2] Katy Juarez,[3] and **Derek R. Lovley**[2] (dlovley@microbio.umass.edu)

[1]University of Tennessee Health Science Center, Memphis; [2]University of Massachusetts, Amherst; and [3]Instituto de Biotecnología/UNAM, Cuernavaca, México

Project Goals: See goals for abstract 149.

Knowledge of how structural genome differences among microorganisms lead to variation in gene regulation is fundamentally important for our understanding of the functioning of gene regulatory pathways and their individual components. This knowledge is also necessary for our better understanding of the genomic changes leading to adaptation to diverse environments. Our research focuses on *Geobacteraceae*, a metal-reducing family of delta-Proteobacteria, which are capable of harvesting electricity from organic matter and

environmental bioremediation of organic and metal pollutants. We are investigating molecular mechanisms which allow these species to adapt and regulate their responses to environmental stimuli which result in energy production and removal of environmental pollutants. In this presentation, we report on our progress in computational analysis of gene regulation in species from this family. Our ongoing analyses include cataloguing and integrating information about gene regulation, *in silico* prediction of transcription factor binding sites, and investigation of functional effects of genome-scale and single nucleotide and amino acid level changes in transcription factors, operons, and regulatory sequence elements in *Geobacteraceae.*

Computational pipeline for analyses of variation in regulatory organization among species and strains. To investigate the regulatory impact of genome-level as well as nucleotide sequence-level evolutionary changes, we have developed a computational pipeline for analysis of regulatory changes in operon organization among species and strains of *Geobacteraceae.* This computational pipeline uses pairwise all-against-all sequence similarity searches for all genes obtained from two genomes of interest. Comparisons are made by MEGABLAST for nucleotide comparisons between strains of the same species and by tblastx for between-species comparisons at the protein sequence level, with soft filtering and E-value thresholds. We employ bi-directional best hits, coordinate matching, and gene neighborhood analyses to identify pairs of orthologs between the two strains and to validate operon assignments in each strain, while simultaneously recording gene annotations within each strain. After ortholog pairs and operon assignments have been unambiguously established, the pipeline tracks categories of structural gene changes (mutations and domain shuffling) and operon changes (e.g., full or partial deletions and duplications of the individual operons, merging of genes into new operons, splitting of individual operons into novel operons, as well as shuffling of groups of genes among different operons). These changes are further compared to known regulon assignments of individual genes of *Geobacter sulfurreducens*, an extensively studied model representative of *Geobacteraceae,* obtained across different microarray data experiments. This genome scale comparison allows identification of correlations between structural genome changes leading to operon shuffling and changes in gene co-regulation. Subsequent steps integrate these changes in genome sequence, gene co-regulation, and single base pair changes with available information on locations of regulatory sequence elements, which are suggested to influence expression of individual genes. We have applied this pipeline to pairwise comparisons of the *G. sulfurreducens* genome with the genome of the hypermutated strain, KN400, adapted for growth on electrodes [1] and with genome sequences of two other species, *G. metallireducens* and *G. uraniireducens.* This approach allowed us to identify operons which merged, split, or underwent gene reshuffling in the process of evolution. We also used bioinformatic approaches to investigate how single nucleotide level changes in the genome of the strain KN400 affected promoters, multiple transcription factor binding sites, and transcriptional and translational attenuators.

Bioinformatic analysis of transcriptional regulation. We are continuing our research investigating transcriptional regulation of specific biological pathways in *Geobacteraceae* which play important roles in electron transfer and environmental response. In our earlier studies, we investigated multiple regulatory pathways controlled by a variety of sigma factors and other transcriptional regulators. Now we are concentrating our analysis on two regulatory systems, the TetR family of transcriptional regulators and an enhancer binding protein, PilR. Members of the TetR bacterial family regulate expression of genes whose products are involved in a variety of important biological functions, e.g., osmotic stress, catabolic pathways, homeostasis, biosynthesis of antibiotics, efflux pumps, and multidrug resistance. Earlier studies of the TetR family [2] suggested that its members may be particularly abundant and diverse in microbial species exposed to environmental changes, due to their likely role in microbial adaptation to environmental changes. Due to their important roles, we investigated molecular evolution and regulation of 9 TetR family members in *G. sulfurreducens* (including functionally important regulators such as OrfR and AcrR, and other regulators with less established roles), their homologs in other species of *Geobacteraceae*, and their more distant relatives in other microbial species. In total, we performed an in depth analysis of 864 TetR family members across species of Bacteria and Archaea, and also investigated conservation of target operons regulated by TetR family members in *Geobacteraceae.* Our analyses demonstrated that TetR family regulators were not uniformly distributed among *Geobacteraceae,* suggesting either gene loss or horizontal gene transfer of some of these genes. While a number of TetR family members have been duplicated in certain species of *Geobacteraceae,* many of these duplications appeared to be species-specific.

We are also continuing our investigation of an enhancer binding protein, PilR, which regulates the expression of the *pilA* gene encoding structural pilin in an RpoN-dependent manner. *G. sulfurreducens* pili are electrically conductive and are required for Fe (III) oxide reduction and for optimal current production in microbial fuel cells. Previously we predicted PilR regulatory sites upstream the *pilA* gene and of other *G. sulfurreducens* genes whose products participate in biosynthesis, assembly, and function of pili and flagella, in secretory pathways, and in cell wall biogenesis. At present time, we are focusing our research on the biological roles of specific targets located downstream of predicted PilR sites, in order to better understand the regulatory role of PilR in *G. sulfurreducens.* In agreement with studies by other research groups, we are observing some intriguing similarities in genome regulatory sites affecting transcriptional regulation of the type II secretion system, which includes *pilA*, among environmentally important species of *Geobacteraceae*, and a number of bacterial pathogens, in which pili are important for pathogenesis and virulence.

Cataloguing regulatory information. We have developed and are continuing to update our online database, GSEL (*Geobacter* Sequence Elements) [3], which compiles regulatory information for *G. sulfurreducens*. At present time, we are incorporating the wealth of novel information on

transcription factor binding sites, attenuator sites, and riboswitches, which has been obtained by multiple research groups using experimental and computational approaches. The GSEL online server, which integrates information on operon organization, gene annotations, and regulatory sequence elements, is continuing to serve as a publicly available resource allowing users to investigate transcriptional regulation and regulatory interactions in *G. sulfurreducens.*

References

1. Yi, H., K.P. Nevin, B.C. Kim, A.E. Franks, A. Klimes, L.M. Tender, and D.R. Lovley, Selection of a variant of Geobacter sulfurreducens with enhanced capacity for current production in microbial fuel cells. Biosens Bioelectron, 2009. 24: 3498-3503.
2. Ramos, J.L., M. Martínez-Bueno, A.J. Molina-Henares, W. Terán, K. Watanabe, X. Zhang, M.T. Gallegos, R. Brennan, and R. Tobes, The TetR family of transcriptional repressors. Microbiol Mol Biol Rev, 2005. 69: 326-356.
3. Qu, Y., P. Brown, J.F. Barbe, M. Puljic, E. Merino, R.M. Adkins, D.R. Lovley, and J. Krushkal, GSEL v. 2, a genome-wide query system of operon organization and regulatory sequence elements of Geobacter sulfurreducens. OMICS, 2009. 413: 439-449.

153

Experimental Annotation of a Bacterial Genome

Yu Qiu[1]* (yuqiu@ucsd.edu), Byung-Kwan Cho,[1] Young Seoub Park,[1] **Derek Lovley**,[2] Bernhard Ø. Palsson,[1] and Karsten Zengler[1]

[1]Dept. of Bioengineering, University of California, San Diego and [2]Dept. of Microbiology, University of Massachusetts, Amherst

Project Goals: See goals for abstract 149.

Genome sequencing has been paramount to our understanding of biology. Precise genome annotation by computational prediction alone, however, has not been accomplished yet. Here we describe an experimental approach to annotate a bacterial genome that integrates proteomics and transcriptomics data with genome-wide data for RNA polymerase and sigma factor binding and single base-pair resolution transcription start site (TSS) determination. The proteogenomics approach alone yielded 55 new ORFs, 36 ORFs in intergenic regions and 19 pORFs in a different frame or on the opposite strand compared to current annotation. Additionally, we confirmed 241 genes that had previously been predicted only as hypothetical proteins, together representing an increase of ~9% compared to current annotation. The integration of transcriptomic data with RNAP holoenzyme binding regions and TSS data resulted in 748 and 694 RNAP-guided transcription segments (i.e. operons) on the forward and reverse strand, respectively, containing 2.3 genes in average. We identified 115 new RNAP-guided transcription segments that have not been annotated before, 70% of those representing antisense transcripts. In addition, the start of over 50 genes was corrected using TSS information and proteomic evidence. Furthermore, a large number of small non-coding RNAs (34), typically difficult to annotate using computational methods, has been identified and validated by Northern blotting. The data presented here demonstrate that more than 10% of the computationally generated annotation can be corrected using experimental data. More fundamentally, not only a structural annotation, i.e. the ORF structure of the genome, has been accomplished but at the same time the genome has been annotated on a higher operational level by elucidating its transcriptional architecture, i.e. the operon structure.

154

De novo Assembly of a Complete Microbial Genome Using Short Reads

Harish Nagarajan[1,2]* (nh@ucsd.edu), Jessica Butler,[3] Anna Klimes,[3] Yu Qiu,[2] Karsten Zengler,[2] Barbara A. Methé,[4] Bernhard Ø. Palsson,[2] **Derek R. Lovley**,[3] and Christian L. Barrett[2]

[1]Bioinformatics and Systems Biology Graduate Program and [2]Dept. of Bioengineering, University of California, San Diego; [3]Dept. of Microbiology, University of Massachusetts, Amherst; and [4]J. Craig Venter Institute, Rockville, Md.

Project Goals: See goals for abstract 149.

The development of next-generation sequencing technologies has greatly reduced the cost of sequencing per base and opened up a wide range of applications. One of the major applications include *de novo* sequencing of new microbial genomes. De novo sequencing using next-generation technologies have necessitated the development of new algorithms for assembling these short and more error-prone reads. Several de novo assembly algorithms (EULER-SR, Velvet, VCAKE, AllPaths, etc.) that are capable of assembling millions of short-reads from next-generation sequencing technologies into thousands of contigs with varying degrees of efficiency, have been recently developed. The complementary nature of Illumina and 454 reads has been exploited by some recent methods that have produced an assembly of *P. syringae pathovar oryzae* which consisted of 126 scaffolds with 2002 unincorporated contigs and an N50 of 91.5 kb. Another report integrated these two data types using a different approach to assemble an *Acinetobacter baylyi* strain into 10 scaffolds with an N50 of 1Mb.

Despite these recent reports that indicate significant progress by integrating Illumina and 454 technologies, complete de novo assembly of microbial genomes from only short reads and without aid from Sanger sequencing still remains an unsolved challenge. This challenge is critically important, for a single, circular nucleotide sequence of the complete chromosome is a necessary prerequisite for confident and complete research based on a genome. This fact is highlighted by a recent commentary that stresses on the importance and applications of high-quality genome sequences.

As an answer to this challenge, we have developed a successful integrative approach (meta-assembly) for combining next-generation sequencing technologies (Illumina and 454) to assemble a complete microbial genome de novo and applied it to a novel *Geobacter* variant (KN400) that is capable of unprecedented current production at an electrode. In addition to integrating the complementary data types (Illumina and 454), our meta-assembly strategy also leverages the different and complementary results provided by multiple assembly programs to obtain the complete sequence of a microbial genome.

Our meta-assembly approach consists of four distinct phases namely, Hybrid Assembly, Scaffold Bridging and Finishing, Scaffold Ordering and Genome Finishing. We applied this strategy using 50X Illumina GA1 singleton reads and 16X 454 GS-FLX paired-end sequencing reads for the novel *Geobacter* variant (KN400). By early integration of the two data types in the Hybrid Assembly phase, we were able to assemble the reads into just a few scaffolds. We then exploited the complementary assemblies generated by Newbler and EULER-SR in order to resolve the degenerate nucleotides in the scaffolds. We employed a PCR-based search strategy in order to obtain the correct relative ordering and orientation of the scaffolds. We corrected for indels and any errors introduced during our scaffold finishing and scaffold ordering phase by aligning the Illumina reads to the ordered scaffold and obtained a circular genome of length 3,714,259 bp.

KN400 is the first complete microbial genome sequence to be assembled from only short reads without the aid of Sanger sequencing. We found the completed KN400 genome to be collinear over its entire length with no major rearrangements and approximately 97% identical at the sequence level, to *Geobacter sulfurreducens* PCA. We further performed a comparative genomics analysis to identify unique genomic regions and thus provide functional insights into the observed novel phenotype of this particular strain.

This readily applicable strategy will result in a significant increase in the number of complete microbial genomes and should impact the quality of current and future sequencing projects. This strategy can also be applied to sequence all unknown members of a microbial community that can be physically separated thus providing the foundation for systems-level characterization of microbial communities by accelerating the rate of obtaining whole genome sequences of each of the community members.

155

Evidence for Direct Cell-to-Cell Electron Transfer from Adaptive Evolution, Genome Resequencing, and Gene Deletion Studies

Zarath M. Summers[1]* (zsummers@microbio.umass.edu), Heather Fogarty,[1] Ching Leang,[1] Ashley Franks,[1] Bernhard Palsson,[2] Yu Qiu,[2] Christian Barrett,[2] Karsten Zengler,[2] and **Derek R. Lovley**[1]

[1]Dept. of Microbiology, University of Massachusetts, Amherst; and [2] University of California, San Diego

http://www.geobacter.org

Project Goals: See goals for abstract 149.

Syntrophic transfer of electrons between cells of different species is a central feature in the function of many anaerobic ecosystems that impact on the global carbon cycle, as well as for several types of bioremediation and production of methane, an important biofuel. For forty years the primary model for interspecies electron transfer has been interspecies hydrogen transfer in which one syntrophic partner disposes of electrons by reducing protons to produce hydrogen and the other partner uses the hydrogen as an electron donor in respiration.

However, interspecies hydrogen transfer is a rather inefficient mechanism for two microorganisms to share electrons because of potential energy loses in the multiple steps required for hydrogen production and reoxidation. We hypothesize that microorganisms can directly exchange electrons via cell to cell electrical contacts that may involve conductive filaments known as microbial nanowires and/or electron transfer via extracellular, multiheme *c*-type cytochromes.

In order to evaluate how syntrophic cell-to-cell electron transfer might evolve a co-culture of *Geobacter metallireducens* and *Geobacter sulfurreducens* was inititated in a medium that contained ethanol as the sole electron donor and fumarate as the sole electron acceptor. Consistent with genome-scale metabolic modeling, *G. metallireducens* can metabolize ethanol, but cannot respire with fumarate and *G. sulfurreducens* can use fumarate as a terminal electron acceptor, but can not metabolize ethanol. However, it is potentially feasible for the two organisms to grow in ethanol-fumarate medium with *G. metallireducens* metabolizing ethanol to acetate and carbon dioxide with the release of electrons and with *G. sulfurreducens* consuming the electrons as well as the acetate with the reduction of fumarate.

Initially the co-culture grew very slowly and required 30 days to metabolize the ethanol provided. However, with continued transfer of a 1% inoculum the co-culture adapted to utilize the ethanol within 3 days. This increase in metabolic rate was accompanied by the formation of large (ca. **x** mm diameter) spherical aggregates. The formation of

aggregates and faster metabolism of ethanol were associated with less release of hydrogen from the co-culture.

The aggregates were comprised of approximately 15% *G. metallireducens* and 85% *G. sulfurreducens*. Fluorescent *in situ* hybridization (FISH) with species-specific probes of thin sections of the aggregates demonstrated that cells of *G. sulfurreducens* completely surrounded small spherical clusters of *G. metallireducens*. Transmission electron microscopy of aggregate thin sections revealed that there were abundant filaments, with the *c*-type cytochrome OmcS attached, coursing between the cells. This is consistent with the potential for direct cell-to-cell electron transfer.

Sequencing of the genomic DNA from the aggregates demonstrated that there was a single-base pair substitution in the gene for PilR in the *G. sulfurreducens* associated with the aggregate. Previous studies in our laboratory have demonstrated that PilR functions as an RpoN-dependent enhancer binding protein and regulates expression of genes encoding pilin and OmcS. Gene knock-in studies are underway to determine whether this mutation is responsible for the apparent pilin-OmcS network between the cells.

In order to further evaluate the role of hydrogen as the currency for cell-to-cell electron exchange a *G. metallireducens-G. sulfurreducens* co-culture was established with a strain of *G. sulfureducens* in which genes for the uptake hydrogenase, HybB, were deleted. This co-culture formed large, spherical aggregates within a matter of weeks, rather than the 7 months that were required for wild-type cells to form aggregates. These results suggest that preventing interspecies hydrogen transfer may increase selective pressure for direct cell-to-cell electron transfer. Genome resequencing of this syntrophic pair is underway to determine whether there were specific mutations associated with the more rapid development of aggregates.

These studies suggest that direct cell-to-cell electron transfer may be feasible and could be the mechanism of choice for cell-to-cell electron transfer when there is intense selective pressure for rapid syntrophic metabolism. Additional studies are underway with a diversity of mutants in which genes for *c*-type cytochromes, other redox active proteins, or pili have been deleted in order to further evaluate this hypothesis.

156

The Application of Metagenomic and Metatranscriptomic Methods to the Study of Microbial Communities in a Uranium-contaminated Subsurface Environment

Barbara Methé[1]* (bmethe@jcvi.org), Robert DeBoy,[1] Michael Geimer,[1] Peter Rosanelli,[1] and **Derek Lovley**[2]

[1]J. Craig Venter Institute, Rockville, Md. and [2]University of Massachusetts, Amherst

Project Goals: See goals for abstract 149.

Here we report on the application of metagenomic and metatranscriptomic approaches to the study of the microbial community at the Old Rifle site located in Rifle, Colorado (a former uranium ore-processing facility which is currently being managed as a part of the Uranium Mill Tailings Remedial Action (UMTRA) program of the U.S. Department of Energy) The study presented here is meant to complement and contribute to the wider ongoing investigations at this site through the application of next generation sequencing to metagenomic and metatranscriptomic analyses in an effort to reveal genetic information from subsurface microorganisms without the need for prior cultivation.

A metagenomic framework for use in this project was established from the 2007 sampling of the Rifle D05 site groundwater. A hybrid assembly has been constructed using 454 FLX pyrosequencing from which ~121MB of sequence was obtained from a 3-4 kb insert paired-end library and ~58 MB worth of Sanger sequencing paired-end reads from a 4-5kb insert plasmid library. Taxonomic classifications of either reads or assemblies of 454 and Sanger data, identified approximately half can be assigned at the Family level as members of the Geobacteraceae. Further, contigs that can be assigned to the genus *Geobacter* and are members of the Subsurface Clade 1, such as *G. bemidjiensis*, *G, uraniumreducens* and *G.* sp M21, predominate. From a whole genome alignment analysis, coverage of ~71% of the *G. bemidjiensis* genome was also revealed.

An incremental clustering method using the Cd-hit algorithm (a fast clustering algorithm based on shared word counts) was applied to the set of ~238,000 ORFs called from reads from both assemblies, and reads unincorporated into assemblies, to identify and classify protein families. Results from this investigation reveal that the most abundant of the clusters at 90% identity, and containing at least 100 members, are a reverse transcriptase family followed by several transposases indicating the ubiquitous nature of mobile elements in the microbial community. However, other highly abundant protein families of bacterial origin identified include DNA-directed RNA polymerase, a c-type cytochrome family, several families whose functions are most likely related to cell to cell, and cell to surface contact, and a family of heavy metal efflux pumps. In addition, several abundant hypothetical protein families were identified. This latter finding underscores one of the strengths of

metagenomic analysis as new protein families can be identified that are not currently in the database. Results from the incremental clustering were also used to produce an estimate of library coverage using a rarefaction analysis which revealed that despite the significant amount of sequencing completed, the library has not been saturated (in other words, additional sequencing would reveal new ORFs and protein families).

The hybrid assembly was further evaluated through the JCVI Metagenomic pipeline which provides functional annotation of ORFs from multiple sources and is used to assign gene names, Enzyme Commission (EC) numbers, gene symbol, GO identifiers, and biological role categories. Further annotation and data mining was also performed using MG-RAST tools. An examination of HMM profiles (TIGRfams and Pfams) revealed that among the most abundant profiles are those with functions related to environmental sensing and signal transduction (e.g., PAS domain proteins, diguanylate cyclase domain proteins and sigma-54 interaction and response receiver domain proteins). Other prevalent HMM profiles based on relative abundance include profiles with functions related to exopolysaccharide sorting or transport to the cell surface and efflux transport.

Results from the annotation of ORFs from assemblies were also used to construct metabolic pathways to better understand the metabolic potential of the microbial community. Based on the use of KEGG maps, >50% overall representation of enzymes (including full and partial pathways) was determined in the categories of Purine and Pyrimidine Metabolism, Energy Metabolism, Amino Acid Metabolism and Carbohydrate Metabolism. Other categories of interest for which full and partial pathways could be constructed were the Biosynthesis of Secondary Metabolites and the Biodegradation of Xenobiotics categories. For instance, complete pathways were determined for 1,2-dichloroethane degradation and 3-chloroacrylic acid degradation. While 60% of the enzymes present in the pathway for benzoate degradation via CoA ligation, and 40% of the enzymes present in the pathway for benzoate degradation via hydroxylation, respectively were identified.

We have also begun experimentation designed to elucidate community gene expression through investigation of the metatranscriptome at several Rifle sites (including the 2007 samplings of D05 and D07 and the 2008 sampling of D04). We have used a capture oligonucleotide approach to enrich mRNA which is then used as template in cDNA reactions for subsequent sequencing via 454 pyrosequencing. A comparison of expression profiles from each site reveals an important result of this approach which is the identification of reads most closely related to multiple hypothetical ORFs and enzymes of unknown function from previously sequenced genomes. These findings suggest that these ORFs are true coding sequence whose biological roles remain to be determined. Further, each expression profile reveals evidence of the presence of non-coding RNAs which serve important roles in gene regulation. A comparison of each site in terms of taxonomic diversity reveals that although relatives to Subsurface Clade 1 organisms predominate in the metatranscriptomes, differences in microbial composition do exist. For example, the highest percentage of reads related to Subsurface Clade 1 organisms (~50%) were identified at the D05 site. A wide variety of transcripts with biological role categories has also been identified at each site and these results are currently undergoing additional investigations. Among the functions with greatest abundance of transcripts determined thus far, are ones related to protein synthesis and environmental sensing and signaling. Overall, the efforts presented here are central to several goals of this project especially the investigation of *in situ* microbial communities to generate new insights and hypotheses concerning metabolic processes and interactions between community members. Further, these results may also serve as a framework for predictive modeling of processes relevant to bioremediation.

Student Presentation

157

Tuning the Conductivity and Capacitance of *Geobacter sulfurreducens* Biofilms by Regulation of Gene Expression

Nikhil S. Malvankar[1,2]* (nikhil@physics.umass.edu), Kelly P. Nevin,[2] Ashley Franks,[2] Byoung-Chan Kim,[2] Kengo Inoue,[2] Tünde Mester,[2] Mark T. Tuominen,[1] and **Derek R. Lovley**[2]

[1]Dept. of Physics and [2]Dept. of Microbiology, University of Massachusetts, Amherst

http://www.geobacter.org

Project Goals: See goals for abstract 149.

The surprising finding that microorganisms can transfer electrons over substantial (>100 μm) distances to the surface of electrodes has expanded the potential applications of microbe-electrode interactions. Optimization of these applications requires an understanding of the mechanisms for this long-range electron transfer. One proposed mechanism is that, under the appropriate conditions, microorganisms may be able to form electrically conductive biofilms. However, previous studies have found that microbial biofilms act as insulators rather than conductors.

Geobacter sulfurreducens produces the highest current densities of any known microorganism. In order to investigate the possible conductivity of *G. sulfurreducens* biofilms, cells were grown in a microbial fuel cell that contained two gold electrodes as a "split-anode", separated by 50 μm non-conductive gap. *G. sulfurreducens* grew on the split anode forming a biofilm that bridged the non-conductive gap. Both the dc current-voltage characteristics and the ac impedance response between the two gold electrodes demonstrated that *G. sulfurreducens* biofilms are highly conductive. Various controls revealed that biofilm conductivity depended upon the physiological status of the cells.

To evaluate components that might confer conductivity to biofilms, novel electrode-selected variants and mutants, deficient in different outer membrane *c*-type cytochromes, were grown in the split-anode device. Strains generating more current produced biofilms with higher conductivity, demonstrating that higher biofilm conductivity facilitates higher current density and suggesting that conduction is an important mechanism for long-range electron transfer through the biofilms. Western immunoblot analysis revealed a direct correspondence between biofilm conductivity and the abundance of pilin protein. This is significant because previous studies have suggested that these pili are electrically conductive and may function as 'microbial nanowires'. An electrochemical gating technique further suggested the direct role of pili in biofilm conductance.

There was no correspondence between the number of *c*-type cytochromes and conductivity. Furthermore, biofilm conductance remained unchanged even after the activity of the cytochromes was inhibited by treating the biofilm with a reagent that unfolded the cytochromes. However, there was a direct relationship between capacitance and the abundance of cytochrome hemes in the biofilms. Capacitance decreased two orders of magnitude after cytochromes were treated with a denaturing reagent, further demonstrating that cytochromes contribute to capacitance. This is consistent with the concept that the abundant multi-heme *c*-type cytochromes in the periplasm and outer membrane of *Geobacter* species play an important ecological role, permitting temporary storage of electrons and allowing continued short-term electron transfer across the inner membrane when natural electron acceptors are temporarily unavailable.

These results demonstrate for the first time that biofilms can be electrically conductive and suggest that pili play an important role in conductance. The potential to grow conductive films and alter their electronic properties via genetic engineering is a significant advancement in the emerging field of bioelectronics.

158

CRISPR-Mediated Inhibition of Histidyl-tRNA Synthetase in *Geobacter sulfurreducens* Recreates an Evolutionary Pressure Experienced by *Pelobacter carbinolicus*

Muktak A. Aklujkar* (muktak@microbio.umass.edu) and **Derek R. Lovley**

University of Massachusetts, Amherst

http://www.geobacter.org

Project Goals:
Overall Goal: The primary goal of this research is to develop experimental and computational tools to predictively model the behavior of complex microbial communities involved in microbial processes of interest to the Department of Energy.

Five Year Goal: The five year goal is to provide computational tools to predictively model the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for *in situ* bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. Based on our current rate of progress, it is expected that within five years it will be possible to predict the *in situ* growth and activity of the relevant microorganisms in the environments of interest solely from relevant geochemical data and to describe in detail the *in situ* metabolic state of the microorganisms from environmental gene expression data. Furthermore, these computational tools will be able to predict the response of the microbial community to environmental manipulations or manipulation of the genome of the relevant organisms, allowing rational optimization of *in situ* uranium bioremediation or electricity harvesting via environmental or genetic engineering.

The family *Geobacteraceae* primarily contains Fe (III)-respiring bacteria capable of uranium bioremediation and electricity production. *Pelobacter carbinolicus*, a member of the family *Geobacteraceae*, cannot reduce Fe (III) directly, or produce electricity, and lacks the abundant *c*-type cytochromes implicated in electron transfer to insoluble extracellular acceptors. The evolutionary basis for these differences is an intriguing problem, and the sequencing of several genomes of *Geobacteraceae* allows it to be addressed. Genome analysis of *P. carbinolicus* revealed a sequence of 32 base pairs (spacer #1) within the clustered regularly interspaced short palindromic repeats (CRISPR) locus, which is identical to a segment of the histidyl-tRNA synthetase (*hisS*) gene. Recent studies have established CRISPR loci as the immunological memory of microbes, with individual spacers matching genes within phage or plasmid entities that previously infected the host, and providing resistance against recurring infection. However, spacers that match host housekeeping genes have not received attention. The model of CRISPR function predicts that spacer #1 should inhibit expression of *hisS* in *P. carbinolicus*. A shortage of histidyl-tRNA, in turn, should impede translation of proteins with multiple closely spaced histidines, predisposing them to mutation or elimination from the genome. A combination of genomics and genetics was used to investigate whether *P. carbinolicus* experienced this evolutionary pressure. Comparison of the *P. carbinolicus* genome with those of four other *Geobacteraceae* confirmed that genes with high histidine demand (computed as the number of histidines divided by the harmonic mean distance between histidines) are fewer in *P. carbinolicus* than in *Desulfuromonas acetoxidans*, *Geobacter bemidjiensis*, *Geobacter metallireducens* and *Geobacter sulfurreducens*. Sixteen gene families that are clearly ancestral have been either mutated to reduce histidine demand or lost by *P. carbinolicus*: these include several *c*-type cytochromes and a subunit of NADH dehydrogenase. Thus, inhibition of *hisS* by spacer #1 could have caused the metabolism of *P. carbinolicus* to shift away from respiration of extracellular electron acceptors. Moreover, when the *hisS* gene of *Geobacter sulfurreducens* was replaced with that of *P. carbinolicus*, growth with Fe

(III) as electron acceptor was totally inhibited by spacer #1 in the context of the *G. sulfurreducens* CRISPR, and growth with fumarate was strongly inhibited. The amount of *hisS* transcript RNA was reduced in this strain compared to a control lacking spacer #1, but the amount of transcript for a control gene, *hisZ*, was similarly reduced, indicating that this effect may be due to the reduced growth rate rather than RNA-level inhibition. Over forty transfers, the inhibited strain adapted to grow as well as the control, and to grow on Fe (III), but spacer #1 was not eliminated. The genome of the adapted strain was resequenced, and appears to bear a deletion of a noncoding region within a cluster of CRISPR-associated genes. Together, these investigations demonstrate that evolution of *P. carbinolicus* from a *Geobacter*-like ancestor, including the loss of multiheme cytochromes, which have high histidine demand, can be attributed in part to CRISPR-mediated inhibition of histidyl-tRNA synthetase. This is the first successful co-introduction of a CRISPR spacer and its target in the same cell, the first application of a hybrid CRISPR construct consisting of a spacer from one species in the context of repeats of another species, and the first report of a potential impact of CRISPR on genome-scale evolution by inhibition of an essential enzyme's expression.

Program Title: Genome-Based Models to Optimize *In Situ* Bioremediation of Uranium and Harvesting Electrical Energy from Waste Organic Matter

159

Genome-Scale Reconstruction of the Bacterial Transcriptional Regulatory Networks

Byung-Kwan Cho[1]* (bcho@ucsd.edu), Eric M. Knight,[1] Young Seoub Park,[1] **Derek Lovley**,[2] and Bernhard O. Palsson[1]

[1]Dept. of Bioengineering, University of California, San Diego; and [2]Dept. of Microbiology, University of Massachusetts, Amherst

Project Goals: See below.

The tremendous amount of novel genomic information is inspiring a new understanding of the bacterial genome on a global scale. Interactions between proteins and DNA for example are fundamental for cellular functions such as transcription, DNA replication, recombination, chromosome packing, and DNA repair. Among these, transcription is an essential step in gene expression and its understanding has been one of the major interests in molecular and cellular biology. By precisely tuning gene expression, transcriptional regulation determines the molecular machinery for developmental plasticity, homeostasis, and adaption.

The primary goals of this project are to develop genome-based experimental approaches to elucidate the function of anaerobic microbial communities involved in groundwater bioremediation and novel bioenergy processes and to develop genome-scale models of these communities in order to better understand and optimize bioremediation and bioenergy applications. It is expected that this study will provide experimental and computational tools that can be used to predict the response of the microbial community to environmental manipulations, or manipulation of the genome of the relevant organisms, allowing rational optimization bioremediation and bioenergy applications via environmental or genetic engineering. To address the issues, the genome-wide transcription factor binding regions (TFBR) - where transcription factors (TFs) bind to modulate the binding of the RNA polymerase (RNAP) - for the global transcription factors in *Escherichia coli* as a model system has been determined using ChIP-chip assays using high-resolution whole genome-tiling microarrays. Our results identified unique and reproducible TFBR from exponentially growing cells under various growth conditions. Most of the TFBR are located within intergenic regions; however significant number of TFBR were found within open reading frames and between divergently transcribed regions.

Changes in the binding levels of TFBR in response to environmental stimuli were strongly correlated with the changes in levels of mRNA transcript and RNAP occupancy. Three distinct regulatory modes (independent, concerted, and reciprocal modes) were identified through the integration of the ChIP-chip, mRNA transcript level, and RNAP occupancy level data. For the further understanding of the regulatory modes, we determined the alternative uses of transcription start sites (TSS) using massive-scale sequencing and subsequently reconstructed the causal relationships between TFBR and the TSS at the genome scale. Previously, we observed that ~35% of promoter regions contain multiple TSSs indicating the presence of alternative TSSs for large portions of the *E. coli* transcription units. Our results identified over several hundreds of direct interactions between global transcription factors (i.e., Fnr, ArcA, Crp, FruR, and Lrp) and TSS under the respective growth conditions. For example, there are two TSSs (3,595,753 and 3,595,778) for the *livKHMGF* operon. Although the TSS (3,595,753) is dominantly used to transcribe the operon, the transcription factor Lrp represses the other TSS (3,595,778) under exponentially grown cells in the absence of leucine. As demonstrated here, the alternative TSSs are used to regulate the bacterial transcriptome in response to different environmental stimuli. The phenomenon is likely to be tightly linked with the transcription regulatory network. Also, the use of alternative TSSs might be widespread in bacteria including the subsurface microbial community.

160

Periplasmic/Extra-Cytoplasmic Sensor Domains of Microbial Two-Component Transmembrane Signal Transduction and Chemotaxis Proteins

P. Raj Pokkuluri,[1] Hendrik Szurmant,[2] Andrzej Joachimiak,[1] and **Marianne Schiffer**[1]* (mschiffer@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Argonne, Ill. and [2]Dept. of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, Calif.

http://www.bio.anl.gov/structural_biology/correlates.html

Project Goals: As sub-project of GTL grant "Genome-based models to optimize in situ bioremediation of uranium and harvesting electrical energy from waste organic matter, Derek Lovley (PI)" our goals are to analyze selected proteins to understand their function in the cell. This includes modeling of structures based on their amino acid sequences, determination of their structures, and the functional interpretation of the structures, such as active sites and surface properties.

Transmembrane signal transduction in bacteria is primarily mediated by the so-called two-component signal transduction system. Some of these proteins have extra-cytoplasmic sensor domains that interact with the environment of the bacteria and trigger responses in the cytoplasmic domains that effect the functioning of the cell. Chemotaxis proteins (Mcp) are part of this system that help the bacteria to sense its environment and move to a more favorable location. To generate an appropriate response to an environmental signal, a sensor histidine kinase (HisKa) adjusts its intrinsic autokinase activity upon detection of ligand(s). To understand the signals that trigger these proteins we have studied their soluble periplasmic sensor domains of these membrane proteins. Elucidating of the stimulatory signals and the corresponding transduction of the message by the above multidomain proteins is relevant in bacteria of both environmental and health significance.

The genome of *Geobacter sulfurreducens*, a Gram-negative organism important for bioremediation, encodes for a large number of signal transduction/chemotaxis proteins. 82 HisKa, 14 with peirplasmic sensor domains; 32 Mcp, 21 with periplasmic sensor domains; 28 GGDEF, 4 with periplasmic sensor domains. We previously determined the structures of two periplasmic sensor domains of chemotaxis receptor proteins that contained a *c*-type heme from *G. sulfureducens* (1) and found that they adopted a PAS-like fold as we had predicted (2). The structure determination by X-ray crystallography for both proteins revealed a novel PAS-like structure formed by two protein chains that involved domain swapping (1). This is the first time that a PAS-like structure was found for a Mcp sensor domain and not a four helix bundle structure found in the aspartate receptor, Tar.

Recently, we are focusing on periplasmic sensor domains of signal transduction proteins from another Gram-negative bacterium of high importance for bioremediation, *Anaeromyxobacter dehalogenans*. *A. dehalogenans* also has a large number of two-component signal transduction molecules: 76 HisKa, 23 with peirplasmic sensor domains; 18 Mcp, 8 with periplasmic sensor domains; 10 GGDEF, none with periplasmic sensor domains. The predicted sensor domains are in the Midwest Center for Structural Genomics (MCSG) pipe line for cloning, protein expression and structure determination.

In an extension of our effort to understand the structure-function of sensor domains of signal transduction proteins in general, we worked with the gram-positive model organism, *Bacillus subtilis*. We used structure-prediction programs to predict that the extra-cytoplasmic sensor domains of HisKa proteins of *B. subtilis* and found that 11 out of 13 have a PAS-like fold (3). This prediction was verified by the determination of the structures of two of these domains as part of the MCSG effort. The determined sensor domain structures showed that our predictions were correct. Because of the large sequence variability of the PAS-like domains are not identified by sequence searches, but we found that the PAS-like fold is a common structural module used by bacteria for detection of its environment (3).

References

1. P.R. Pokkuluri, M. Pessanha, Y.Y. Londer, S.J. Wood, N.E.C. Duke, R. Wilton, T. Catarino, C.A. Salgueiro, M. Schiffer. Structures and Solution Properties of Two Novel Periplasmic Sensor Domains with *c*-Type Heme from Chemotaxis Proteins of *Geobacter sulfurreducens*: Implications for Signal Transduction. *J. Mol. Biol*. 377, 1498-1517 (2008).
2. Y.Y.Londer, I. S. Dementieva, C. A. D'Ausilio, P. R. Pokkuluri, M. Schiffer. Characterization of a *c*-type heme containing PAS sensor domain from *Geobacter sulfurreducens* representing a novel family of periplasmic sensors in *Geobacteraceae* and other bacteria. *FEMS Microbiol. Lett*, 258, 173-181 (2006).
3. C. Chang, C. Tesar, M. Gu, G. Babnigg, A. Joachimiak, P. R. Pokkuluri, H. Szurmant, and M. Schiffer. Extra-cytoplasmic PAS-like Domains are Common In Signal Transduction Proteins. J. Bacteriology, in press.

161

Systems-Level c-di-GMP Signaling in *Shewanella oneidensis* MR-1

Lily Chao, Shauna Rakshe, and **Alfred M. Spormann***
(spormann@stanford.edu)

Depts. of Chemical Engineering, and of Civil and Environmental Engineering, Stanford University, Palo Alto, Calif.

Project Goals: The objective of this research as part of the *Shewanella* Federation, has been to develop a systems-level understanding of c-di-GMP signaling. C-di-GMP

has been recognized as an emerging intracellular prokaryotic signaling nucleotide with perhaps greater implications than c-AMP, and is currently being intensely studied.

Cyclic di-GMP (c-di-GMP) is an important intracellular prokaryotic signaling molecule that is involved in regulating a wide variety of pathways including metabolism, cell division, motility, and pathogenesis. The environmental and cellular factors controlling c-di-GMP signaling are numerous and diverse, but are not well understood. Diguanylate cyclases (DGC) characterized by a 'GGDEF' domain and c-di-GMP-specific phosphodiesterases (PDE) characterized by an 'EAL' domain are known to alter intracellular c-di-GMP concentrations. However, the mechanisms by which these enzymes regulate the cell in response to environmental and cellular conditions are not well understood. In *Shewanella oneidensis* MR-1, many of these enzymes also contain a sensor domain such as the Per-Arnt-Sim (PAS) domain, which can respond to changes in redox potential, oxygen, other small molecular ligands, or light, as well as facilitate protein-protein interactions. Here, we present physiological and biochemical data comparing several PAS-GGDEF-EAL domain proteins present in *Shewanella oneidensis* MR-1. One of these proteins, SO0341, appears to control amino acid metabolism through c-di-GMP signaling. Growth analyses of the deletion mutant in planktonic cultures indicated an extended lag phase compared to that of wild type. Interestingly, growth could be rescued by supplementation of the media, specifically with the amino acids isoleucine, leucine, and valine. Transcriptional microarray analyses of the SO0341 deletion mutant showed that expression of many genes involved in the isoleucine, leucine, and valine biosynthetic pathways are reduced. Enzymatic activity assays of the SO0341 protein indicate that it has both DGC and PDE activity. We have also identified two flavin-binding PAS-GGDEF-EAL proteins, SO0141 and SO3389. Deletion mutants of SO0141 and SO3389 exhibit similar phenotypes in which both strains are less motile under anaerobic growth conditions. Enzymatic activity assays demonstrated that both proteins exhibit DGC and PDE activity. Hence the flavin in these enzymes may be involved in sensing the oxygen or redox status of the environment and modifying the activity of the enzyme accordingly. In contrast to SO0141 and SO3389, a deletion mutant of SO0437 is less motile and exhibits larger biofilms only under aerobic conditions. Furthermore, this phenotype is only apparent under rich medium growth. Our results demonstrate that these enzymes regulate a variety of molecular functions including central metabolism (SO0341) as well as motility (SO0141, SO3389, SO0437).

162

Identification of Genes Involved in Anaerobic Sulfite Reduction by *Shewanella oneidensis* MR-1

Sheetal Shirodkar and **Daad Saffarini*** (daads@uwm.edu)

University of Wisconsin, Milwaukee

Project Goals: One of our goals is to identify components of the electron transport chain that lead to the reduction of sulfur compounds by *S. oneidensis* MR-1. In addition, we aim to identify regulatory proteins that control the reduction of these compounds and elucidate their role in anaerobic respiration.

Shewanella oneidensis MR-1 is a metal reducer that uses a large number of electron acceptors for respiration, including sulfur compounds. Reduction of thiosulfate, polysulfide, and tetrathionate is accomplished using PsrABC that is similar to thiosufate reductases from other bacteria. The mechanism of sulfite reduction by *S. oneidensis* MR-1, however, is poorly understood. We have identified an operon that encodes genes required for sulfite reduction. These include c cytochrome maturation genes required for heme ligation to atypical heme-binding sites. The operon also encodes a c cytochrome and proteins involved in copper transport. Additionally, we have determined components of the electron transport pathway that are needed for sulfite reduction. These include c cytochromes, such as CymA and menaquinones. Analysis of molybdopterin cofactor biosynthesis mutants indicated that sulfite reduction requires a molybdopterin-containing protein. We have also identified a transcriptional regulator, encoded by SO_0490 that regulates sulfite reduction. Mutants deficient in SO_0490 are unable to reduce sulfite. The effect of SO_0490 on the transcription of the genes involved in sulfite reduction, described above, was tested using reverse transcription PCR. The results indicated that SO_0490 does not regulate the expression of these genes. This suggests that additional components are involved in sulfite reduction by *S. oneidensis*. Experiments are underway to identify these proteins and heir role in this process.

163

Investigating Hydrogen Ecology in Marine Microbial Mats Through Integration of Biogeochemistry, (Meta)Transcriptomics and Nano-Scale Imaging

Luke Burow[1,2]* (lburow@stanford.edu), Dagmar Woebken,[1,2] Tori Hoehler,[2] Brad Bebout,[2] Alfred M. Spormann,[1] **Lee Prufert-Bebout**,[2] **Steven W. Singer**,[4] **Jennifer Pett-Ridge**,[3] and **Peter K. Weber**[3]

[1]Dept. of Chemical Engineering and Civil and Environmental, Stanford University, Stanford, Calif.; [2]NASA Ames Research Center, Moffett Field, Calif.; [3]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.; and [4]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: The goals of this project are (1) to develop the capability to link identity and function in complex microbial communities and (2) characterize the hydrogen ecology of microbial mats.

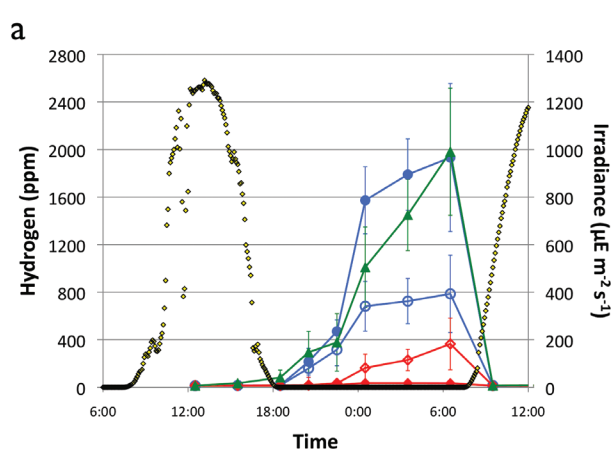
Microbial mats are diverse communities of microorganisms that can produce or consume hydrogen (H_2). Proposed industrial-scale production of biohydrogen and other biofuels will inevitably grow 'unwanted' biofuel-consuming organisms that negatively impact upon biofuel production. Thus, these mat systems represent a valuable model for investigating interactions between microbes that positively and negatively affect biofuel production. We are developing a combination of biogeochemical, molecular and imaging techniques to characterize microbial mat hydrogen ecology.

We have observed H_2 and organic acids to be produced in the upper 2-3 mm layer of mats from Elkhorn Slough, California at concentrations that are orders of magnitude above day levels. Manipulation of the photoperiod by either depriving the mat of light or inhibiting oxygenic photosynthesis using 3- (3,4-dichlorophenyl)-1,1-dimethylurea (DCMU) prior to the H_2-producing dark period indicated that oxygenic phototrophs are likely to be important H_2-producers (Fig. 1a). Spatial localization of H_2-producers and H_2-consumers are a critical determinant of net H_2 production in these systems as disruption of the mat architecture by mechanical homogenization increased H_2 production substantially (Fig. 1a). Addition of molybdate to inhibit sulfate-reducing bacteria (SRB) increased net hydrogen production (Fig. 1a, b) and reduced hydrogen sulfide production suggesting that SRBs are important H_2-consumers in the mats.

Suppression of N_2-fixation using excess ammonium did not decrease net H_2 (Fig. 1b) or organic acid production suggesting that fermentation is not stimulated by a requirement for energetically expensive N_2-fixation and that the main source of H_2 in microbial mats is due to fermentative H_2 production. Although H_2 and organic acid production decreased when the daytime photoperiod was withheld from the mats, it was still substantially higher during the night than during the day. These observations suggest that fermentation is regulated by a circadian rhythm.

Metatranscriptomics approaches are capable of identifying a broad range of genes expressed *in situ* and are exploited in this investigation with a view to capturing key catabolic genes, including fermentative genes related to hydrogen producing pathways. A subtractive rRNA approach using probes complementary to 16S, 23S, 18S and 28S rRNA was used to reduce the high percentage of rRNA often obtained from total RNA isolated from environmental samples. Our method is an adaptation of that currently commercially available from Ambion Inc. (MICROB*Express*). The probe mixture included probes that specifically target cyanobacterial and lower eukaryotic ribosomal sequences in addition to probes already available in the kit. Application of a comprehensive suite of probes reduced rRNA sequences from ~90% to ~40%. In addition to a global metatranscriptomic approach we are using a directed approach that specifically identifies phototrophic and heterotrophic bidirectional hydrogenase (H_2-ase) genes via H_2-ase clone libraries and H_2-ase microarrays. These strategies will provide diversity information of H_2-ases present, which will infer the phylogenetic identity of H_2-producers and/or consumers. Additionally these strategies will be used to evaluate temporal H_2-ase expression dynamics.

The biogeochemical and molecular data we have gathered is enabling nanometer-scale functional and phylogenic imaging by NanoSIMS (NanoSIP-FISH) to identify key organisms involved in fermentation and hydrogen cycling. Daylight incubations with ^{13}C-bicarbonate combined with fluorescence in situ hybridization is being used to identify organisms that fix carbon and generate photosynthate, which is the proposed source of electrons for H_2 evolution. Performing NanoSIP-FISH with samples incubated with ^{13}C-acetate under dark, anoxic conditions will label sulfate-reducing bacteria that consume acetate and hydrogen.



* Presenting author

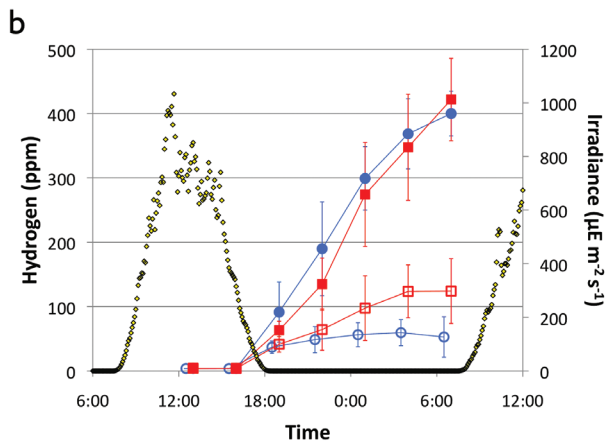


Fig. 1. Net hydrogen production in Elkhorn Slough microbial mats. (a) Effect of photoperiod deprivation (red filled diamonds), 20 µM DCMU (red unfilled diamonds), 30 mM molybdate (filled blue circles) and homogenization (green filled triangles) compared to a control (unfilled blue circles) on H_2 production. (b) Effect of N_2-fixation suppression on H_2 production under control and SRB inhibited conditions. 8.8 mM ammonium added to suppress N_2-fixation (red squares) and compared to N_2-fixing incubations (blue circles). Filled red squares and filled blue circles signify incubations with 30 mM molybdate.

164

NanoSIP: Combining Stable Isotope Probing and High Resolution Secondary Ion Mass Spectrometry to Identify Diazotrophs in Complex, Stratified Microbial Communities

Dagmar Woebken[1,2]* (dwoebeken@stanford.edu), Luke Burow,[1,2] Lee Prufert-Bebout,[2] Brad Bebout,[2] Tori Hoehler,[2] Jennifer Pett-Ridge,[3] Steven W. Singer,[4] Alfred M. Spormann,[1] and **Peter K. Weber**[3]

[1]Dept. of Chemical Engineering and Civil and Environmental, Stanford University, Stanford, Calif.; [2]NASA Ames Research Center, Moffett Field, Calif.; [3]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore,Calif.; [4]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: See below.

Understanding the metabolic processes performed by complex microbial communities requires new methods to observe nutrient fluxes *in situ* and to link the identity and function of microbial community members. Cyanobacterial mats are complex, stratified microbial communities that are often found in coastal marine environments and governed by the diel cycle. During the day, oxygenic photosynthesis accumulates fixed carbon, which is fermented during the night under anoxic conditions. Under these anoxic conditions, these mats evolve large amounts of H_2 as a byproduct of fermentation and N_2 fixation. Therefore, the mats are an excellent model system to understand ecological factors that affect how solar energy is captured, stored in chemical form and then released as a potential biofuel. To link the H_2 ecology of these mats to the flux of carbon and nitrogen, we are developing a new technique, NanoSIP, which combines stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) to measure nutrient uptake and assimilation at the single cell level. Previous studies of similar cyanobacterial mats indicated N_2 fixation at high levels and that diazotrophic microbial communities in these mats are comprised of a diverse range of filamentous cyanobacteria, unicellular cyanobacteria and heterotrophic bacteria. The application of ^{15}N-NanoSIP to these mats will allow us to estimate the contributions of individual taxa to the N_2 fixation observed in the sample and correlate ^{15}N incorporation in single cells with *nifH* expression. Additionally, we will be able to correlate carbon and nitrogen flux in samples treated with $^{13}HCO_3^-$ and $^{15}N_2$.

To demonstrate this method, we are studying H_2-evolving cyanobacterial mats collected at Elkhorn Slough in central California. These mats receive substantial inputs of fixed nitrogen from agricultural runoff during the rain season, yet can fix N_2 at high rates during most of the year, suggesting that diazotrophy is an important component of the nitrogen cycle in these mats. Using the acetylene reduction assay, we have identified mat samples dominated by *Microcoleus* spp. that fix N_2 at night and, unexpectedly, mat samples with numerous benthic heterocystous cyanobacteria that fix N_2 primarily during the day. Incubation of these mat samples with $^{15}N_2$ and subsequent analysis by isotope ratio mass spectrometry (IRMS) demonstrated ^{15}N incorporation into biomass. The majority of the label was incorporated into the upper 2mm of the mats, the "green layer" where cyanobacteria are localized. These ^{15}N-labeled samples have been analyzed by NanoSIMS to determine which microbes fix $^{15}N_2$ (Figure 1). Diazotrophic microbes will be identified by linking ^{15}N label observed by NanoSIMS measurements with cell morphology and element labeling using CARD-FISH (Catalyzed Reporter Deposition-Fluorescence *in situ* Hybridization). We use this technique, referred to as EL-FISH, to identify diazotrophic unicellular cyanobacteria and heterotrophic bacteria, which cannot be distinguished by morphology. To complement these single cell techniques we are identifying expressed *nifH* genes recovered from cDNA clone libraries and extracted from metatranscriptomic datasets.

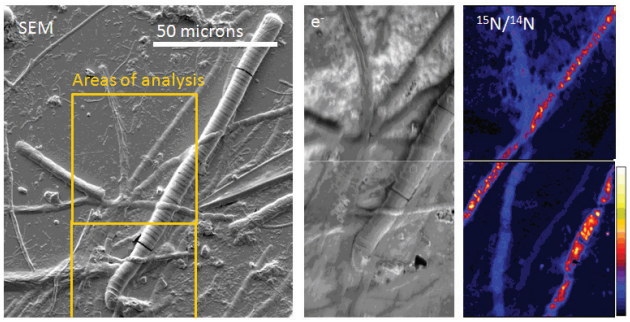


Fig. 1. Scanning electron microscopy (SEM) micrograph and NanoSIMS secondary electron (e-) and nitrogen isotope ratio ($^{15}N/^{14}N$) tiled images of the cyanobacterial community of the Elkhorn Slough microbial mat. The mat was incubated in $^{15}N_2$ for 9 hours during the night prior to sampling. Cyanobacteria that were fixing nitrogen incorporated the ^{15}N tracer. The color scale bar represents $^{15}N/^{14}N$ ratios of 0.003 (black) to 0.09 (white).

165

Population Genomics Reveals a Genetic Basis for Ecological Differentiation in Two Subpopulations of Ocean Bacteria

B. Jesse Shapiro, Jonathan Friedman, Otto X. Cordero, Sarah Preheim, Dana Hunt, Sonia Timberlake, **Martin Polz**, and **Eric Alm*** (ejalm@mit.edu)

Dept. of Civil and Environmental Engineering and Biological Engineering, Massachusetts Institute of Technology, Cambridge

Project Goals: Understand genomic structure, horizontal transfer, and natural diversity in *Prochlorococcus* and *Vibrio* strains

http://proportal.mit.edu/
http://almlab.mit.edu/
http://web.mit.edu/polz/lab/home.html

Microbes adapt to changing selective pressures in their natural environments, leading to a potentially dynamic process of ecological specialization and speciation, even over short periods of time. Yet little is known about the micro-evolutionary processes leading to ecological differentiation of microbial populations in the wild. We sequenced and analyzed complete genomes from 8 closely-related strains of *Vibrio splendidus*, representing two nascent populations, that appear to have recently diversified ecologically: 3 strains found primarily on small particles, and 5 strains found primarily attached to zooplankton. Although gene-flow between populations in the two habitats is common, we observe a significant excess of recent recombination within habitats, suggesting the emergence of ecologically differentiated populations. Gain and loss of DNA is extensive among these strains (each strain contains ~100-300 kb of strain-specific DNA), and a few recently aquired genes may provide habitat-specific adaptive value. For example, a suite of genes involved in O-antigen and mannose-sensitive hemagglutinin (MSHA) biosynthesis are absent in small-particle strains but present in zooplankton-associated strains, perhaps promoting preferential attachment to zooplankton. We also identified a few 'core' genomic regions that, while present in all strains, are highly differentiated between the two habitats. These ecologically associated loci include genes involved in stress response (rpoS), DNA repair (cysteine methyltransferase), and chitin metabolism, leading us to hypothesize that switching between zooplankton-associated (rich in insoluble exoskeletal chitin) and small-particle-associated lifestyles may require fine-tuning of chitin metabolism.

166

Transcriptional-Translational Offsets in the Diel Cell Cycle of *Prochlorococcus*

Jacob R. Waldbauer,[1]* Sébastien Rodrigue,[2] and **Sallie W. Chisholm**[2] (chisholm@mit.edu)

[1]Joint Program in Chemical Oceanography, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution, Cambridge; [2]Dept. of Civil and Environmental Engineering and Dept. of Biology, Massachusetts Institute of Technology, Cambridge

Project Goals: Identify posttranscriptional regulatory processes governing the diel cell cycle of *Prochlorococcus*; link biogeochemical processes to cellular metabolism through quantitative proteomics.

The marine cyanobacterium *Prochlorococcus* is the most abundant oxygenic photosynthetic organism on earth and a key component of oceanic carbon and nutrient cycling. In the ocean, the growth of *Prochlorococcus* cells is tightly coupled to the diel light/dark cycle, and this growth cycle is responsible for primary production on the order of 10 gigatons of carbon per year. Experiments in culture have shown that the mRNA-level expression profiles of most *Prochlorococcus* genes have a strong diel rhythm. We quantified the abundance and periodicity of the *Prochlorococcus* proteome over a diel cycle and compared transcript-level and protein-level expression patterns. Strong diel oscillations in transcript abundance are broadly damped at the protein level, and temporal offsets between the two suggest that posttranslational regulatory mechanisms are important in determining the abundance dynamics of a number of proteins. The overall composition of the proteome is quite stable over the diel cycle, with some proteins consistently among the most abundant in the cell at all times of day, despite major differences in cellular metabolism between light and dark periods. The small variations in protein abundance that accompany significant changes in central metabolic activities imply that *Prochlorococcus* biochemical networks may be poised near balance points that allow for redirection of metabolic fluxes with relatively small shifts in the abundance of their components.

167

Cyanophages Encode and Express the Calvin Cycle Inhibitor CP12 and Pentose Phosphate Pathway Enzymes

Luke R. Thompson, Qinglu Zeng* (qinglu@mit.edu), Libusha Kelly, Katherine Huang, Maureen L. Coleman, and **Sallie W. Chisholm**

Dept. of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge

http://chisholmlab.mit.edu/

Project Goals: Characterize the cellular machinery of *Prochlorococcus* and its viruses as a model system for photosynthetic energy conversion.

Marine cyanophages are known to carry and express genes for the light reactions of photosynthesis. Viral photosynthesis proteins are thought to boost host photosynthesis and help repair damaged photosystems. Notably, no Calvin cycle genes have been reported in cyanophages, suggesting that the products of photosynthetic electron transport are used directly and not stored as glucose. Instead, cyanophages carry genes for the pentose phosphate pathway (PPP), a pathway for glucose oxidation. In one case, the PPP enzyme transaldolase is regulated to be co-expressed with photosynthesis genes, implying that photosynthesis and the PPP operate concurrently in infected cells. These observations led us to the following hypothesis: electron flow through photosynthesis is critical for a productive cyanophage infection, but that energy is not stored as glucose; rather, energy from photosynthesis and from glucose oxidized by the PPP is used to power phage replication. To address this hypothesis, we searched all available cyanophage genomes—and metagenomic sequence of likely viral origin—for any evidence of Calvin cycle and PPP genes. We also measured gene expression of key metabolic genes during infection to address the coordination of transcription. Although no Calvin cycle enzymes were detected, we found widespread incidence of the Calvin cycle inhibitor CP12, particularly in T4-like myoviruses. In cyanobacteria, CP12 binds and inhibits two enzymes in the Calvin cycle, acting as a metabolic switch to direct carbon flux away from the Calvin cycle and toward the pentose phosphate pathway. Three PPP genes were also widely distributed in cyanophages: *zwf* (glucose-6-phosphate dehydrogenase), *gnd* (6-phosphogluconate dehydrogenase), and *talC* (transaldolase). *talC* and *cp12* were the most prevalent of these four genes in cyanophage genomes, followed by *gnd* and *zwf*, and this hierarchy was mirrored in metagenomic databases. PPP, photosynthesis, and DNA biosynthesis genes were co-expressed with known T4-like early genes during myovirus Syn9 infection of *Synechococcus* WH8109. Thus, phage-encoded proteins for all three pathways may play a role early in infection, working in concert. The presence of CP12 and PPP genes and absence of Calvin cycle genes in cyanophages collectively suggests that phage-augmented photosynthesis is used for energy production but not carbon fixation during infection. Rather, glucose is likely oxidized by the PPP, and the NADPH and ribose generated may be used for phage nucleotide production and to relieve oxidative stress induced by infection.

168

Unlocking the Illumina Genetic Analyzer Platform for Improved De Novo Sequencing and Metagenomics

Sébastien Rodrigue* (s_rod@mit.edu), Arne Materna, Sonia Timberlake, Rex R. Malmstrom, **Eric J. Alm**, and **Sallie W. Chisholm**

Dept. of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge

http://chisholmlab.mit.edu/
http://almlab.mit.edu/

Project Goals: Obtain longer sequencing reads from the Illumina Genetic Analyzer platform in order to expand its range of applications to de novo genome assembly and metagenomics while significantly reducing the cost of DNA sequencing projects.

DNA sequencing is one of the most powerful tools for accessing the genetic and metabolic diversity found in microorganisms. The new generation of sequencing instruments allow massive amounts of information to be obtained at an unprecedented speed. Three platforms, the Roche-454 Genome Sequencer, the Illumina Genetic Analyzer, and the Applied Bio-System SOLiD system are currently widely available. The latter two instruments produce a very large number of short reads at a significantly lower cost per basepair. However, their limited readlength creates notorious difficulities in applications such as de novo sequence assembly and metagenomics. Strategies to obtain longer sequencing reads from the Illumina Genetic Analyzer or ABI SOLiD instruments would allow to further exploit their potential, expand their range of applications, and in some cases, significantly reduce the cost associated with DNA sequencing projects.

We describe a simplified procedure to prepare high-throughput sequencing libraries, as well as a strategy to obtain sequencing reads that average 150-210 basepairs long with the Illumina Genetic Analyzer II platform. Following an automatable gel-less library construction, two converging reads are obtained and overlapped to reconstitute a longer composite read. We show that approximately 90% of the paired-end reads can successfully be assembled with this strategy. The approach allowed de novo assembly of single-cell amplified genomes from *Prochlorococcus* cells isolated from the south Pacific ocean. We also demonstrate that the method can be applied to metagenomics sequencing, producing significantly more data at a more affordable cost.

169

Microbial Ecology and Genomics Across Natural Gradients of Light, Nutrients, and Carbon Flow

C.L. Dupont[1]* (cdupont@jcvi.org), Z. Johnson,[2] R.K. Stuart,[1] J. Badger,[1] J. McCrow,[1] L. Zeigler,[1] K. Barbeau,[1] B. Palenik,[3] and **A.E. Allen**[1]

[1]J. Craig Venter Institute, Rockville, Md.; [2]University of Hawaii; and [3]Scripps Institution of Oceanography, La Jolla, Calif.

Project Goals: Compare metagenomic signatures of bacterial and eukaryotic populations of marine microbes across a light and dissolved iron gradient in surface and chlorophyll maximum communities in the Southern California Bight.

The physical, chemical, and biological gradients associated with the transition from the coast to the central Pacific are immense. Offshore of San Diego, water column integrated carbon fixation declines up to 80% within fifteen kilometers. Over the next 700 kilometers, integrated carbon fixation rates fall another one to two orders of magnitude. Nitrate supported primary production, which is likely proportional to carbon export from the euphotic zone, fall a thousand fold over the 700 km. Often thermally stratified, the Southern California Bight also has a subsurface chl a maximum at the nitracline. As such, this area provides a perfect natural laboratory for studying how gradients in light and carbon flow influence microbial community composition and genome content. In the summer of 2007, the gradients in microbial biomass and physiology were characterized over a 700 km oceanographic section leading from coastal San Diego to the edge of the Pacific Gyre. Three distinct oceanographic regimes were identified by their physical and biological characteristics, including light field, depth of thermocline and nitracline, carbon fixation, nitrogen uptake rates, and integrated biomass. For each regime, metagenomic libraries were constructed for both the subsurface chlorophyll maximum and surface microbial communities. Phylogenomic analyses revealed a large enrichment in sequences attributable to viruses and archaea in the subsurface chl max. In general, microbial diversity was higher in the chl max relative to the surface samples. The average bacteria genome size was higher in the subsurface chl max, but also scaled according to cell size and carbon flow. The coastal to open ocean gradient was characterized by an increased abundance in proteins associated with particle adherence, Fe scavenging, and trace metal detoxification.

170

Microbial Metagenomics of the California Current Ecosystem: Southern Upwelling and Northern Oxygen Minimum Zone

Lisa A. Zeigler[1,2]* (lzeigler@jcvi.org), Jonathan H. Badger,[1] John P. McCrow,[1] Ian T. Paulsen,[3] Eric E. Allen,[2] Shannon J. Williamson,[1] and **Andrew E. Allen**[1]

[1]J. Craig Venter Institute, San Diego, Calif.; [2]Scripps Institution of Oceanography, University of California, San Diego; and [3]Macquarie University, Sydney, Australia

Project Goals: To assess microbial community structure and function of the California Current ecosystem using metagenomic sequencing of large volume (200L) water samples. Taxonomic profiles and functional potential are being examined to gain a greater understanding of which organisms are involved in biogeochemical cycling and how they are performing these roles.

Metagenomics has become a powerful tool for enabling the discovery of vast genetic diversity relating to microbial communities; primarily the uncultivated populations. Additionally, the use of metagenomics has led to novel discoveries and new paradigms within marine microbiology. Two datasets taken as a part of the GOS expedition from within the California Current ecosystem are described.

In collaboration with the California Cooperative Oceanic Fisheries Investigations (CalCOFI), which is a partnership between the California Department of Fish and Game, the NOAA Fisheries Service and the Scripps Institution of Oceanography that focuses on the study of the marine environment off the coast of the California, a unique subset of metagenomic samples was collected from the California Current and Southern California Bight (July 2007) to be integrated into the Global Ocean Sampling (GOS) expedition. Exploiting the immense hydrographic metadata acquired by CalCOFI provided a rare opportunity to use metagenomic analyses to describe taxa and metabolism that correspond to the nutrient-rich upwelling waters consistent to this region. These samples are unique to the primarily open ocean sites of the GOS expedition based on their location across a coastal upwelling gradient where carbon, nitrogen, phosphorous and micronutrients are delivered to the pelagic food web. It is our goal to assess the role, through genomics, of different microbial groups in controlling these fluxes. The dataset consists of approx. 4.7Mbp across seven sites from 3 different size ranges, 20-3.0um, 3.0-0.8um, and 0.8-0.1um. These sites span upwelling regions with a relative increase in nutrients and chlorophyll to oligotrophic regions. For taxonomic evaluation of predicted ORFs, we applied a phylogenomic approach using APIS, an automated pipeline that builds trees for every ORF within a dataset. Significant differences in community composition of the microbial populations across the upwelling gradient were revealed. Oligiotrophic sites displayed similar taxa to the GOSI sample dataset, which are typified by a dominance of Alphaproteobacteria and Cyanobacteria, primarily SARII

and *Prochlorococcus* sp., respectively. Four sites representing regions of higher inorganic nutrient availability displayed clear differences in taxa representation with elevated levels of Gammaproteobacteria, Bacteriodetes, and eukaryotic prasinophytes. One of these upwelling stations was characterized by a Planctomycete bloom. Additional to the taxonomic and nutrient correlation, relationships between functional categories, nutrients, and taxa are being distinguished. Analyses of transporter proteins, for example, suggests variation in substrate utilization between the different water masses; specifically oligotrophic sites appear to encode higher numbers of peptides for ammonium and amino acid transport, whereas, upwelling sites are enriched in nitrate MFS (major facilitator superfamily) and ferrous ion transporters. Ongoing analyses are aimed at linking bacterial and eukaryotic populations present, as well as, further annotation of key functional categories.

Secondly, in collaboration with members of the Coastal Margin Observation and Prediction (CMOP), JCVI researchers (L. Zeigler and A. Allen) have been examining the coastal ecosystem off Oregon pertaining to the Columbia River system (CRS). The CRS affects ecological processes in the Northern California Current System and therefore correlates to other datasets currently being investigated at JCVI. Here four samples were taken that cover freshwater influx into the region to assess the microbial communities; 1) CRE – Columbia River estuary 2) CR plume – Columbia River plume (chlorophyll maxima of ocean just outside plume of freshwater influx) 3) OMZ - 80m sample of anoxic region 4) BTM - 1200m sample of bottom ocean layer. As above, using APIS the following initial taxonomic findings are represented. Briefly, the CRE site has an abundance of actinobacterial sequences and the CR plume waters have an increase in Proteobacteria (primarily "*Candidatus Pelagibacter ubique*"), Euryarchaeota, Bacteroidetes, and dsDNA viral sequences. It appears that the CRE bacterial assemblage is comprised of smaller genomes and the gene pool (particularly in the case of the most dominant group, Actinobacteria) is enriched in functions that encode for processes related to organic nitrogen utilization (xanthine uracil permease, oligopeptide amino acid and urea transporters). Bacterial populations with larger genomes, on the other hand, dominate CR plume waters in the chlorophyll max (e.g., Bacteroidetes) and encode a large repertoire of inorganic nitrogen transporters and a seemingly more diverse complement of signaling genes. Community composition of the OMZ shows a shift from Proteobacteria (primarily *C. Pelagibacter*) to the CFB group (Chlorobi, Flavobacteria and Bacteriodetes) (primarily Polaribacter). Initial screening of functional classifications show a marked increase in the OMZ sample of genes encoding carbamyol phosphate synthase (CPS), as opposed to other GOS environmental datasets. Further investigations are underway to complete functional analyses of this dataset.

171

Phylogenetic Informatics for Indian Ocean Metagenomics

Jonathan H. Badger (jbadger@jcvi.org), John P. McCrow, Chris L. Dupont, Shibu Yooseph, and **Andrew E. Allen**

J. Craig Venter Institute, San Diego, Calif.

http://www.jcvi.org/cms/research/projects/gos

Project Goals: The development of a robust phylogenetic pipeline for the taxonomic binning of metagenomic reads as well as a database and related tools to allow mining of the data in relation to associated metadata such as nutrient levels.

The second phase of the Global Ocean Survey sampling expedition (GOS II), a global circumnavigation survey of ocean surface waters, provided extra metagenomic coverage of sites in the Indian Ocean. These metagenomic reads, which were created by a combination of Sanger and 454-titanium technology, were analyzed by our APIS (Automated Phylogenetic Inference System) pipeline.

APIS uses BLASTP to compare each predicted protein (or protein fragment) against an in-house database of proteins from all completed genomes. The full-length sequences of each matching protein are extracted and aligned, and a bootstrapped neighbor-joining tree of each alignment is created. The trees are automatically analyzed to identify the closest sequence and organism to the query leaf, and summaries over all genes being analyzed are generated.

We have constructed a relational database containing the APIS results as well as available metadata associated with the sampling sites. This database is in the same format as other JCVI metagenomic projects using APIS in order to aid comparison across projects.

Initial analysis of the Indian Ocean data finds that the most prevalent organism by far in the open ocean samples (although it does not comprise a majority) is the SAR11 alphaproteobacterium *Pelagibacter ubique*, an organism known to thrive in low nutrient conditions. Coastal samples show more diversity, as befitting their greater richness of nutrients, and show a greater occurrence of cyanobacteria.

172

Insights and Comparisons of Indian Ocean Metagenomics

John P. McCrow* (jmccrow@jcvi.org), Jonathan Badger, Chris L. DuPont, Shibu Yooseph, and **Andrew E. Allen**

J. Craig Venter Institute, San Diego, Calif.

Project Goals: The development of computational tools to assess and compare marine metagenomic samples on the basis of phylogenetic and functional diversity, and to examine genome level adaptations to resource availability in different size classes of marine bacteria.

http://www.jcvi.org/cms/research/projects/gos

The Global Ocean Survey (GOS) continues to sample and analyze marine microbial life around the world from the Sorcerer II research vessel. Comparisons of microbial populations in the Indian Ocean to those found during GOS phase I have resulted in a better understanding of variation in ecological niches available in marine surface waters.

Improved methods were developed for estimating bacterial genome equivalents from multiple single-copy core genes to normalize genomic samples prior to comparison. Estimation of average bacterial genome size and taxonomic profiling was performed using these same core genes to provide appropriately normalized comparisons between samples. Analyses of gene family distributions in relation to average genome size indicate significant enrichments in transporter abundance in smaller genomes. New methods have been implemented in order to evaluate overall patterns in diversity in relation to system productivity and stability, leveraging independent phylogenetic analysis of all ORFs and of core genes. A subset of sites in the Indian Ocean were selected for metagenomic sequencing of multiple filter sizes of 3.0μm and 0.8μm in addition to the standard 0.1μm filter allowing for comparison between planktonic populations of bacteria and surface attached bacteria unable to pass through a 3.0μm pore.

Several clear trends have emerged; generally, the open ocean, free living or "planktonic" niche is characterized by low overall diversity, small genomes, and disproportionate enrichment for particular genes families such as transporters. Larger size class or attached bacteria and those coming from more productive habitats on the other hand display higher levels of diversity, distinct gene family signatures, and relatively nitrogen rich proteomes.

173

Metagenomic Characterization of Novel *Prochlorococcus* Clades From Iron Depleted Oceanic Regions

Douglas B. Rusch[1]* (drusch@jcvi.org), **Adam Martiny**,[2] **Christopher L. Dupont**,[1] Aaron L. Halpern,[3] and J. Craig Venter[1]

[1]J. Craig Venter Institute, Rockville, Md.; [2]University of California, Davis; and [3]Complete Genomics, Inc., Mountain View, Calif.

Project Goals: We are exploring how marine microbes adapt to variable environmental conditions.

Broadly, we are interested in how marine microbes adapt to different environmental conditions over time and space. This provides insights into the mechanisms driving microbial evolution as well as describing what aspects of the environment are important to life on earth. As part of the Global Ocean Sampling project we have explored microbial life in surface marine waters across the globe. Many of these microbial organisms that dominate this environment are uncultivated and uncharacterized. Identifying and understanding these organisms is important for understanding how these ecosystems behave now and will behave in the future.

Prochlorococcus is the most abundant marine photosynthetic microbe on Earth. It is primarily found in oligotrophic waters across the globe and plays a crucial role in energy and nutrient cycling in the marine ecosystem. The abundance, global distribution, and cultivatability of *Prochlorococcus* have made it a model system for understanding marine microbial diversity and biogeochemical cycling. Analysis of seventy-three metagenomic samples from the Global Ocean Sampling expedition acquired in the Atlantic, Pacific and Indian Oceans revealed the presence of two related but previously unrecognized *Prochlorococcus* clades. A phylogenetic analysis using three different genetic markers places the clades close to known high-light adapted lineages. The two clades consistently co-occur and dominate the surface waters of high temperature, macronutrient replete, low iron regions of the Eastern Pacific Equatorial upwelling and the northern edge of Indian Ocean gyre. These new clades are genetically distinct from each other and other high-light *Prochlorococcus* ecotypes and define a novel ecotype. A detailed genomic analysis indicates that the cells from these clades have adapted to iron-depleted environments by reducing their iron quota through the loss of several iron-containing proteins that likely function as electron sinks in the photosynthetic pathway in other high-light *Prochlorococcus* ecotypes. The presence and inferred physiology of these novel clades may explain why *Prochlorococcus* populations from iron-deplete regions do not respond to iron fertilization experiments and further expand our understanding of how phytoplankton adapt to variations in nutrient availability.

174

Responses of Soil Microbial Communities to Long Term Elevated CO_2 in Six Terrestrial Ecosystems

Cheryl R. Kuske[1]* (Kuske@lanl.gov), John Dunbar,[1] Gary Xie,[1] Larry Ticknor,[1] Stephanie Eichorst,[1] La Verne Gallegos-Graves,[1] Shannon Silva,[1] Carolyn Weber,[1] Don Zak,[2] Rytas Vilgalys,[3] Chris Schadt,[4] Dave Evans,[5] Patrick Megonigal,[6] Bruce Hungate,[7] Rob Jackson,[3] Andrea Porras-Alfaro,[8] and Susannah Tringe[9]

[1]Bioscience Divison, Los Alamos National Laboratory, Los Alamos, N.M.; [2]University of Michigan, Ann Arbor; [3]Duke University, Durham, N.C.; [4]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [5]Washington State University, Pullman; [6]Smithsonian Environmental Research Center, Edgewater, Md.; [7]Northern Arizona University, Flagstaff; [8]Western Illinois University, Macomb; and [9]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: Our three goals are to: (1) Understand the impacts of long-term elevated CO_2 and other environmental factors (ozone, nitrogen interactions) on the structure and activities of soil microbial communities, at the DOE's six Free Air CO_2 Enrichment (FACE) and Open Top Chamber (OTC) experimental field sites. (2) Establish a multi-tier genomics-based analysis and ecological integration capability that links DOE JGI sequencing technology with the ability to understand functional abilities of soil microbial communities in an ecological setting. (3) Improve our basis for comparison of soil populations involved in carbon cycling and climate change response by expanding the functional genes and proteins we can use to detect and monitor these populations.

Increased plant growth in response to elevated atmospheric CO_2 results in increased carbon inputs to the soil. The collective activities of the complex soil microflora determine whether this additional carbon is sequestered in the soil or released back into the atmosphere. The contributions of soil microbiota to carbon cycling have local, regional and global impacts on terrestrial carbon storage and cycling. However, our current understanding of the composition and activities of microbial biomass, the major processes that represent control points in carbon flux, and the rates at which they occur in terrestrial ecosystems is rudimentary. Accurate climate modeling and carbon management scenarios require an understanding of these soil processes.

For the past ten years, the DOE has operated six large, replicated field experiments (FACE and OTC experiments) designed to test the effects of elevated CO_2 and other factors on terrestrial ecosystems. Our ability to conduct rigorous field comparisons at the 10 yr endpoint of these experiments, and build upon the existing metadata from the sites, provides an unparalleled opportunity to define critical parameters in soil response to climate change variables. For analysis, we have established a collection of replicate soil samples at each of six ecosystem types encompassed by the DOE's FACE and OTC research sites. In some cases, we also have obtained samples across seasons and in multiple years.

Using targeted (rDNA and functional genes) and shotgun metagenomic sequencing strategies, and quantitative PCR methods, we are studying the impacts of over ten years of elevated CO_2 on the soil microbial communities across these six different ecosystems. It is clear from prior research at these sites that soil microflora have responded to the climate change factors and that the populations and mechanisms underlying those responses are likely to be complex. Across the sites, we are attempting to determine if soil community responses are due to changes in total biomass or biomass of major microbial groups, changes in community composition from growth and/or inhibition of specific populations, or changes in respiration and metabolic activity with no change in total biomass or community composition.

Assessment of the bacterial communities in soils across the sites has shown that the bacterial communities have responded to elevated CO_2 in some ecosystems but not in others, and that the nature of the response is specific to that ecosystem. Responses have included changes in relative abundance of the bacterial community relative to the fungal community, as well as changes in community richness and composition. Where multiple factors were included in the field site experiment (e.g. soil depth, plant species, ozone treatment), the observed community responses associated with those factors were often larger than the direct response to elevated CO_2. A similar assessment of the fungal communities has identified very different fungal populations that dominate the different ecosystems, and changes in relative abundance are correlated with elevated CO_2 at some of the sites. In order to facilitate fungal community comparisons, we have developed a new naïve Bayesian classifier to bin fungal LSU sequences into reliable, validated taxonomic units, and are establishing a large sequence training set for fungal LSU sequences. These tools will be made available to the scientific community through a collaboration with the Ribosomal Database Project (Jim Cole, lead).

175

Exploration of Soil Microbial Cellulose Degradation Through a Synthesis of Enzymatic and Targeted Metagenomic Approaches

Cheryl R. Kuske,[1] **Andrew Bradbury**,[1] **Stephanie A. Eichorst**[1]* (seichorst@lanl.gov), Carolyn Weber,[1] Sara D'Angelo,[1] La Verne Gallegos-Graves,[1] Nileena Velappan,[1] Csaba Kiss,[1] Gary Xie,[1] Terri Porter,[2] and Rytas Vilgalys[2]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, N.M. and [2]Duke University, Durham, N.C.

Project Goals: The overall goals are to (a) identify novel cellulolytic enzymes in specific soil bacteria and in soil bacterial metagenome DNA that can be used as molecular markers in field studies, (b) identify dominant, active bacterial and fungal species that metabolize cellulose across different soil types, (c) identify active cellulolytic fungal enzymes in soils under different climate change scenarios.

Through culture-based studies, it is firmly established that many species of soil fungi and bacteria can use complex plant carbon substrates for growth. However, we still know very little about the microbial species capable of using complex plant carbon in soils or the mechanisms through which these substrates are collectively decomposed in the soil. We are studying bacterial and fungal cellulolytic communities through three related experimental efforts. The overall goals are to (a) identify novel cellulolytic enzymes in specific soil bacteria and in soil bacterial metagenome DNA that can be used as molecular markers in field studies, (b) identify dominant, active bacterial and fungal species that metabolize cellulose across different soil types, (c) identify active cellulolytic fungal enzymes in soils under different climate change scenarios.

(a) We are using a phage-display folding-reporter approach to identify novel cellulose-binding domains (CBDs) of proteins in the bacterium, *Clostridium thermocellum*, and in soil DNA enriched in bacterial species capable of degrading cellulose. This approach can be considered to be the protein equivalent of shotgun sequencing, and could be termed the "domainome". The *C. thermocellum* genome is such a rich source of cellulase genes that it provides an excellent model system to experimentally identify novel cellulases or other enzymatic activities linked to CBDs. We have created appropriate display vectors, and have identified suitable cellulose antigens for selection and DNA sources. Large libraries of protein domains from *C. thermocellum* are in the process of being selected against cellulose for the identification of cellulose binding domains. To apply this technique more directly to discovery of cellulolytic enzymes in soil metagenomes, a large phage display library is being generated from a ^{13}C-enriched bacterial DNA fraction from one of the soils in our stable isotope probing experiments (described in the next paragraph).

(b) Using a combination of stable isotope probing and metagenomic analysis of soil microbial communities, we have identified dominant bacterial and fungal species that actively degrade ^{13}C-cellulose in microcosms containing soils that differ in physical and chemical characteristics as well as background organic matter composition. Our results have identified members of the bacterial phyla *Burkholderia* and *Acidobacteria,* and the fungal genera *Arthrobotrys*, *Chaetomium*, and *Trichocladium* as dominant members of ^{13}C-cellulose degrading populations in soils. ^{13}C-enriched DNA fractions were also enriched in a fungal cellobiohydrolase gene. These studies will allow us to correlate the active and dominant microbial species with soil parameters that will guide future field experiments.

(c) Soil fungi are potentially important in mediating carbon cycling in forest soils, both as plant mycorrhizal associates and as primary decomposers of plant litter and belowground plant biomass. We are investigating the combined influence of long term elevated CO_2 and nitrogen fertilization on the biomass and activities of the soil fungal community through gene expression studies of target genes involved in cellulolytic and lignolytic processes and through comparisons of soil metatranscriptomes.

Collectively these studies should expand our knowledge of bacterial and fungal enzymes and genes involved in cellulose degradation in soils, as well as identify species that could dominate these processes in terrestrial carbon cycling.

* Presenting author