DOE/SC-0140

Sections:
Systems Biology Strategies and Technologies for Understanding Microbes, Plants, and Communities
Analytical Strategies for the Study of Plants, Microbes, and Microbial Communities
Biological Systems Interactions
Plant-Microbe Interfaces
The Predictive Microbial Biology Consortium

U.S. DEPARTMENT OF ENERGY
Office of Science

Joint Meeting 2011

Genomic Science Awardee Meeting IX

and

USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Meeting

[Revised: April 14, 2011]

Crystal City, Virginia
April 10-13, 2011

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Germantown, MD 20874-1290

http://genomicscience.energy.gov

Prepared by
Biological and Environmental Research Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

Systems Biology Strategies and Technologies for Understanding Microbes, Plants, and Communities

Analytical Strategies for the Study of Plants, Microbes, and Microbial Communities

183

Proteomics and Pan-omics Driven Analysis of Microbial Communities, Comparative Biology and Environmental Symbiosis

Mary S. Lipton[1]* (mary.lipton@pnl.gov), Stephen J. Callister,[1] Kristin E. Burnum,[1] Roslyn N. Brown,[1] Kim K. Hixson,[2] Janani Shutthanandan,[1] Samuel O. Purvine,[2] Angela D. Norbeck,[1] Matthew E. Monroe,[1] Carrie D. Nicora,[1] Gordon A. Anderson,[1] and **Richard D. Smith**[1] (PI)
Collaborators: Steve Giovannoni,[3] Cameron Currie,[4] Michael Kahn,[5] Donald Bryant,[6] Norman Lewis,[5] and David Kramer[7]

[1]Biological Sciences Division; Pacific Northwest National Laboratory; [2]Environmental Molecular Sciences Laboratory; Pacific Northwest National Laboratory; [3]Oregon State University; [4]University of Wisconsin; [5]Washington State University; [6]Pennsylvania State University; and [7]Michigan State University

Project Goals: This project employs comprehensive global and directed pan-omics analyses and novel informatics approaches (developed in parallel in this program) of microbes, plants and microbial communities to enhance scientific understanding through elucidation of phenotypic relationships between environmentally important microorganisms, characterization of higher organisms, characterization of the metabolic activities within microbial communities, and identification of post-translationally modified proteins.

Inherent to exploiting microbial function for carbon cycling, bioremediation or biofuel production or utilizing plants as energy precursors is the detailed understanding of the physiology of the cell. Cellular functions are dictated by the proteins expressed in the cell, their resident lifetime in the cell, their localization and their modification state. Additionally, these processes mitigate the metabolites in the cell that serve as energy and carbon currency. This project exploits the technological and informatics advances of the pan-omics measurement pipeline at PNNL (as described in the poster by Anderson et al and Metz et al) to address organism-specific scientific objectives developed in conjunction with biological experts for a number of different microbes and plants. In our poster, we highlight the ability to use pan-omics data for, characterization of microbial communities, elucidation of phenotypic and genotyptic relationships, advances in the characterization of protein modification state, and the determination of protein localization in stem, root and leaf tissues of *Arabidopsis*.

Microbes do not live in isolation; therefore, understanding the function of a microbe in the environment and the effect of the environment upon the microbe requires the characterization of the community as a whole. Research on individual microbes takes on a larger significance if the findings about an organism in cell culture can be extrapolated to the activities of the organism within the natural community in the environment. SAR11, also known as *Pelagibacter ubique*, is the dominant heterotrophic bacterial clade in the oceans, where roughly 25% of the 16S rRNA gene sequences retrieved from uncultured marine bacteria belong to the SAR11 group. Evolutionary selection to minimize genome size in large, nutrient-limited ocean populations, known as genome streamlining, has been implicated as an important factor in the evolution of SAR11. These cells have dispensed with many pathways and transporters that are typically present in bacteria with more complex genomes. We have investigated how *Pelagibacter* respond to iron limitation by applying differential measurements using our new pan-omics platform to *Pelagibacter* cell cultures.

The fungus-growing ant– microbe symbiosis is a paradigmatic example of organic complexity generated through symbiotic association and, over the last decade, it has become a model system for studying symbiosis. We have demonstrated an in-depth profiling of the fungal garden complete with bacteria (fungus alone, isolated bacteria, and fungal garden intact) to understand the relationship between the fungus and the bacterial protectors. Proteomics and metabolomic studies of the secreted proteins from the bacteria have been characterized in an effort to understand the relationship between the ants and the fungus. These studies demonstrate, the ability to use pan-omics measurements on an ecosystem level, spanning bacteria to multi-cellular organisms.

The genotype of an organism is the full hereditary genome, while the phenotype is the actual observed biochemical characteristics of an organism. Although the genome of an organism influences its phenotype, phylogenetically diverse organisms can share a common phenotype. As such, genome-based comparisons are limited in describing these common mechanisms in diverse organisms. We have developed a number of proteomics databases for each of the six bacterial phyla known to contain chlorophyll-based phototrophs, including the recently discovered Candidatus *Chloroacidobacteria thermophilum*, which is currently the only known phototroph within the phylum *Acidobacteria*. Using these databases, to investigate diverse bacteria that share a similar photosynthetic phenotype while having vastly different genotypes. With a better understanding of the photosynthetic pathways, and especially the pathways occurring within cyanobacteria and chloroplasts, systems biology approaches will be poised to determine how these organisms can be used to create alternative fuels, as well as their role in the carbon cycle.

Symbiosis is the long-term interaction between different biological species. One sort of symbiosis is nitrogen fixation occurring in specialized symbiotic interactions between plants and bacteria is a major source of useful nitrogen. We have used Pan-omics measurements to elucidate the interaction between *Medicago truncatula* and *Sinorhizobium meliloti*, mapping proteins expressed in both the plant and the symbiont with preliminary understanding of their interaction.

Additional information and supplementary material can be found at the PNNL proteomics website at http://oberproteomics.pnl.gov/

Acknowledgements: This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Portions of this research were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's Office of Biological and Environmental Research Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

184

Proteomics and Pan-omics Measurements for Comprehensive Systems Characterization of Biological Systems

Thomas O. Metz* (thomas.metz@pnl.gov), Scott R. Kronewitter, Qibin Zhang, Aaron T. Wright, Ronald J. Moore, Tao Liu, Weijun Qian, Erin S. Baker, Therese R.W. Clauss, Karl K. Weitz, Carrie D. Nicora, Heather M. Brewer, Daniel J. Orton, Anil K. Shukla, Rui Zhao, Feng Yang, Roslyn Brown, **Joshua N. Adkins, Gordon A. Anderson, Mary S. Lipton,** and **Richard D. Smith (PI)**

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Wash.

http://www.pnl.gov/

Project Goals: This project endeavors to implement and apply advanced capabilities that aim at comprehensive molecular characterization of biological systems, including the extension of proteomics to cover cover post-translational protein modifications and the implementation of broad metabolomics, lipidomic and glycomic measurements. Together with more widely available genomics and transcriptomics capabilities, this project will provide the transformative "pan-omics" measurement capabilities needed to elucidate interacting networks of genes, proteins, and biochemical reactions in biological systems.

The goal of BER's Genome Science Program (GSP) is to achieve a predictive systems level understanding of plants, microbes and biological communities via the integration of fundamental science and technology developments to enable biological solutions to challenges in energy, environment and climate. Achieving this goal requires comprehensive proteomics, metabolomics, lipidomics, and glycomics, i.e. pan-omics, measurement capabilities and the integration of data generated by these approaches. This project aims to facilitate understandings of biological systems by providing pan-omics molecular measurement capabilities that will be applied in biology-driven collaborative projects led by investigators actively engaged in developing systems biology approaches in support of BER's research agenda. Our strategy benefits from advances in high resolution nano-liquid chromatography (LC) separations combined with high mass accuracy mass spectrometry (MS) measurements and other developments that afford large gains in performance and throughput. These efforts also include the automation of key steps in proteomics sample processing; fractionation of protein samples based on surface membrane protein enrichment and subcellular fractionation methods using differential gradient centrifugation; and implementation of novel methods for protein extraction from environmental (e.g. soil) samples. Additional advancements involve the implementation of targeted proteomics methods (e.g. activity-based protein profiling and multiple reaction monitoring) and approaches for the elucidation of protein isoforms (e.g. integrated top-down and bottom-up proteomics) and post-translational modifications (e.g. phosphoproteomics and characterization of protein glycosites).

To facilitate these goals, this project includes efforts to develop and apply new measurement platforms and integrated analytical strategies implemented in concert with the computational advances necessary for handling increased data production rates, improved data processing algorithms, the development of methods to integrate multiple pan-omics data streams, and efforts needed to effectively disseminate results and information to collaborators and the broader scientific community. Developments are driven by and applied in the context of external collaborative projects aimed at garnering the knowledge needed to lay a foundation for predicting behaviors of and manipulating biological systems critical to DOE missions.

Acknowledgements: This research is supported by the Office of Biological and Environmental Research (OBER) of the U.S. Department of Energy (DOE). Portions of this research were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's OBER. Pacific Northwest National Laboratory is operated for the DOE by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

185

Pan-omics Measurements Platform and Informatics Analysis Pipeline

Gordon A. Anderson* (gordo@pnl.gov), Ronald J. Moore, **Joshua N. Adkins,** David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Kevin Crowell, Stephen J. Callister, Therese R.W. Clauss, Kim K. Hixson, Gary R. Kiebel, Brian L. LaMarche, **Mary S. Lipton,** Da Meng, **Thomas O. Metz,** Matthew E. Monroe, Heather M. Brewer, Carrie D. Nicora, Angela D. Norbeck, Daniel Lopez-Ferrer, Daniel J. Orton, Ljiljana Paša-Tolić, David C. Prior, Samuel O. Purvine, John D. Sandoval, Anuj Shah, Yufeng Shen, Anil K. Shukla, Mudita Singhal, Gordon W. Slysz, Aleksey V. Tolmachev, Nikola Tolić, Karl Weitz, Aaron Wright, Rui Zhang, Rui Zhao, and **Richard D. Smith** (rds@pnl.gov)

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Wash.

This project is developing pan-omics measurement and integrative informatics capabilities to enable comprehensive global molecular characterization to understand, model, and potentially manipulate biological systems. These new analytical capabilities are achieved through the application of advanced separations-MS measurement platforms that greatly increase measurement quality and throughput. This new platform combines fast, multidimensional separations (i.e., fast LC in conjunction with millisecond-scale ion mobility separations) with ultra-fast and accurate mass measurement time-of-flight MS, and provides greatly expanded proteome coverage and greater sensitivity, in addition to at least an order-of-magnitude increase in throughput. Pan-omics measurement capabilities are based on essentially identical separations-MS measurement platforms and similar data processing/informatics pipelines; metabolomics, lipidomics, and glycomics measurements, as well as the expanded proteomics measurements. A key element of pan-omics analysis is informatics methodologies to integrate data from various measurements and incorporate approaches for managing and communicating data, data quality, and ambiguities (e.g., the confidence in peptide and protein identifications, modification sites, abundance levels, etc.).

Advanced measurement, informatics, and computational technologies and approaches are being explored and evaluated for possible broader implementation based on their robustness and suitability for implementation in high-throughput pan-omics, their ability to improve data quality, and their potential to facilitate new biological insights from collaborative applications. For example, we have further developed and applied new IMS-TOF MS measurement platforms that greatly extend our current measurement capabilities by providing data production rates an order of magnitude greater than current (e.g., Orbitrap or FTICR MS-based) platforms in addition to significantly enhancing data quality. Application of the new platforms in conjunction with integrated analytical strategies and increasing automation will provide the throughput needed to more routinely and more extensively cover the range of post-translationally modified proteins, as well as other pan-omics measurements.

Development and application of new measurement capabilities and computational tools are essential for generating, processing, integrating, and disseminating data and information from GSP studies of responses over multiple scales that provide a foundation for manipulating biological systems. We are leveraging the extensive experience and capabilities developed to date within the high throughput proteomics facility at PNNL to extend this capability to multiple omics measurements and provide a framework for effective data integration. Also crucial is the ability to manage, integrate, disseminate, and extract information from increasingly large and complex datasets. The measurement advances noted above require corresponding computational and informatics advances necessary for: managing the resulting increased data production rates; evaluating and controlling data quality; processing and integrating data from the various analysis streams; and disseminating data and information to collaborators, users, and the broader scientific community. Thus, we are developing a suite of data analysis tools, data consolidation applications, and statistical packages, as well as visualization software for data interpretation, and that will support integration of the enhanced proteomics and metabolomics data sets. This framework further supports integration of genomics data from public repositories and aim to provide the needed infrastructure to interoperate with the GTL Knowledgebase.

This poster highlights several developments that enable pan-omics measurements. These developments include; 1) Advances in measurement capabilities, 2) Data management and enhanced informatics analysis capability, and 3) Initial developments of an integrative analysis framework.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

186

Informatics Infrastructure to Enable Pan-omics Measurements of Biological Systems

Matthew E. Monroe* (matthew.monroe@pnl.gov), **Joshua N. Adkins, Gordon A. Anderson,** Kenneth J. Auberry, Kevin Crowell, David A. Clark, Gary R. Kiebel, **Mary S. Lipton, Thomas O. Metz,** Ronald J. Moore, Angela D. Norbeck, Daniel J. Orton, Samuel H. Payne, Samuel O. Purvine, John D. Sandoval, Anuj R. Shah, Gordon W. Slysz, and **Richard D. Smith (**rds@pnl.gov)

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Wash.

Overview: We have established a robust, flexible computational architecture and infrastructure to manage the storage and tracking of raw and processed data associated with pan-omics research. This architecture additionally provides a framework to support data integration that will enable the creation of comprehensive pan-omics measurement datasets. The infrastructure has a web-based interface for accessing and updating information and has mechanisms for exporting and processing the data associated with the identified peptides, proteins, metabolites, etc. The informatics infrastructure will continue to evolve and expand to support the advanced measurement platforms and analysis capabilities needed to enable pan-omics studies.

The Pan-Omics Research Information Storage and Management System (PRISM) provides a flexible and robust infrastructure that serves as the foundation for developing data integration workflows for pan-omics data. The existing informatics infrastructure and analysis tools provide the scalable architecture and base capabilities that allow rapid development to enable integration of pan-omics data. The PRISM infrastructure has evolved and expanded since its inception, and its current design allows for continued expansion, including supporting the increased production rates afforded by a new LC-ion mobility mass spectrometer platform. The PRISM architecture employs a "plug-and-play" paradigm, where individual steps of the informatics pipeline are developed as configurable, cohesive, and independent modules that can be readily chained together in multiple ways to create effectively new informatics pipelines. PRISM incorporates community-developed analysis software, commercial software, and a variety of in-house developed tools for peptide ID, deisotoping, quantitation, and data handling. We anticipate adding additional analysis tools developed for new informatics analysis pipelines related to lipidomics, activity-based proteomics, phosphoproteomics, top-down proteomics, metabolomics, etc. The modular nature of PRISM allows us to offload computationally intensive processing tasks to high-performance computing or cloud computing resources that are becoming available to the scientific community (e.g., the Magellan project, Amazon, or the planned GTL Knowledgebase).

The PRISM infrastructure supports a wide array of functions, including tracking research projects and their associated biological samples, managing the storage and tracking of raw and processed data files, and automated software processing of pan-omics data. As research projects become larger and more diverse, we can further expand LIMS-type capabilities supported by PRISM (e.g., tracking instrument operation and maintenance details). PRISM enables facility staff to better plan and define the sample processing and analysis strategy, including the ability to specify sample run batching and blocking parameters, and to annotate samples with processing factors for use in later data analyses.

PRISM provides several interfaces for accessing and exporting both the raw and processed data. The primary portal for interfacing with PRISM is the DMS website, which allows researchers to browse and search existing information, add new information, and export data. PRISM also includes programmatic interfaces to allow batch export of data using standalone software. The Multi Dataset Analysis and Rollup Tool (MDART) provides a mechanism for collaboration, standardization, and scientific documentation of processed pan-omics data. This tool interfaces with PRISM to allow researchers to export, process, and 'filter' data, and provides the flexibility essential for dealing with a variety of data types, application interests, and data analysis needs. MDART uses workflows to define a systematic and repeatable, yet flexible approach to processing data.

The Mass and Time Tag System (MTS) component of PRISM is responsible for collating peptide search results to form accurate mass and time (AMT) tag databases that can be used for high throughput quantitative studies. MTS is federated across a compute cluster, allowing for ready expansion to support the increasing volumes of data that will be generated in pan-omics studies, including supporting the new measurement platform. VIPER and MultiAlign are used to characterize detected LC-MS or LC-IMS-MS features and match those features to the AMT tag databases. These tools now use the Statistical Tools for AMT tag Confidence (STAC) algorithm to assign confidence values to peptides identified via the peak matching process, thus allowing researchers to filter the results to obtain a specified false discovery rate (FDR).

The PRISM system provides a flexible and robust infrastructure that serves as the foundation for developing data integration workflows for pan-omics data. Workflows are a set of connected operations similar to the work of a researcher (e.g., the integration of multiple time points into a time course dataset). As pan-omics measurements and infrastructure expand, we expect to incorporate new workflows to support pan-omics data integration.

This poster will illustrate the current PRISM capability and developments in progress to enable pan-omics data integration from multiple omics analysis workflows as well as to deal with the significant increase in data volumes generated by the new LC-ion mobility measurement platform.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC05-76RLO 1830.

187

Fidelity and Dynamics of DNA Methylation in Plants

Qin Yao,[1] Changjun Liu,[1] John Shanklin,[1] Chuan He,[2] and **John Dunn**[1]* (jdunn@bnl.gov)

[1]Department of Biology, Brookhaven National Laboratory, Upton, N.Y.; and [2]Department of Chemistry and Institute for Biophysical Dynamics, University of Chicago, Ill.

Project Goals: The first goal of this project is to develop defined in vitro systems for determining the mechanisms by which cytosine DNA methylation is normally maintained in plants and how factors leading to DNA oxidative damage impact the fidelity of DNA methylation. A second goal is to develop, demonstrate, and validate a streamlined "DNA target-enrichment" method coupled to bisulfite sequencing for in-depth methylation mapping of specific plant gene sets and their associated control elements.

Epigenetics is defined as the study of heritable changes to genome structure and function that do not change the nucleotide sequence of the DNA. Methylation of cytosine to form 5-mC in genomic DNA is an important epigenetic marker that plays a critical role in regulation of gene expression, chromatin structure and genome stability. In all organisms, cytosine methylation is a postreplicative process. Newly synthesized DNA strands are unmethylated, thus creating hemi-methylated duplexes at replication forks. In both mammals and plants most methylation occurs at the DNA dinucleotide CpG, where both cytosines in the complementary strands of adjacent base pairs are methylated. In mammals the UHRF1 protein recognizes these hemi-methylated CpG sites via its SET-and RING-associated (SRA) domain. Structural studies have shown that the 5-mC residue in hemimethylated DNA bound to UHRF1 is flipped outside of the DNA helix into a specific 5-mC-binding pocket within the SRA domain. This causes DNA looping and allows the N-terminal region of the SRA domain to interact with the DNA's major and minor grooves. The residue requiring modification is then flipped out of the helix and presented to the DNA methyl transferase DNMT1 for addition of the methyl group.

In *Arabidopsis* VIM1 encodes an SRA domain methylcytosine-binding protein that probably functions similarly to UHRF1 in playing a major role in maintaining DNA methylation patterns following DNA replication. To gain further insight into how VIM1 functions, we have cloned and expressed and purified VIM1 and are using electrophoretic mobility-shift and fluorescence anisotropy titration assays to study its interaction with model duplex DNAs containing cytosine or 5-mC in one or both strands.

A fine-scale mapping tool "Bisulfite Patch PCR"[1] is also being used to discover the potential epigenetic regulation underlying the spatial and temporal expression of 26 genes/ transcription factors involved in lignin biosynthesis in *Arabidopsis*. This new approach allows us to process ~100 genes from multiple samples at the same time.

Recent studies of genomic DNA from human brain, neurons and mouse embryonic stem cells have demonstrated that these DNAs contain a sixth base, oxidized 5-mC or 5- hydroxymethylcytosine (5-HmC). Current thinking is that 5-HmC does not merely mimic 5-mC groups but likely plays an independent role in yet unknown epigenetic regulation of various biological processes. We are using bacteriophage T4 β-glucosyltransferase to transfer a glucose moiety containing an azide group onto the hydroxyl group of 5-HmC. The azide group can then be chemically modified with biotin for detection, affinity enrichment and subsequent sequencing of 5-HmC–containing DNA fragments to reveal the genomic locations of 5-HmC in the DNA[2]. Dot blot detection of biotinylated glucose-HmC shows about 0.06% 5-Hmc in plant leaf genomic DNA. Efforts are underway to map these sites. We are also determining if 5-HmC effects VIM1 binding and if 5-HmC residues can be removed by the *Arabidopsis* DNA glycosylase/lyase, Repressor of Silencing, ROS1, thereby allowing the 5-HmC generated by oxidation of 5-mC, to be replaced by C, resulting in active demethylation.

References

1. Varley, K.E. and Mitra, R.D. (2010) Bisulfite Patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. *Genome Res*, **20**, 1279-1287.
2. Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X. et al. (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol, 29, 68-72.

Early Career Program
Speaking Wednesday 10:30 a.m.

188

A Systems Biology, Whole-Genome Association Analysis of the Molecular Regulation of Biomass Growth and Composition in *Populus deltoides*

Matias Kirst* (mkirst@ufl.edu)

School of Forest Resources and Conservation, University of Florida Genetics Institute, Gainesville

Project Goals: This project main goal is to apply an association genetics approach to unveil the molecular basis of biomass productivity and composition. To comprehensively capture the genetic variants that regulate traits of value for bioenergy production, we are combining sequence-capture and high-throughput sequencing to genotype coding and regulatory sequences in the whole-genome of *P. deltoides*. To achieve this goal we have: (1) optimized sequence-capture for unbiased, high-throughput and low-cost recovery of target coding and regulatory sequences in *P. deltoides*. A set of over 220,000 probes that efficiently capture exon and 500 bp of putative regulatory sequences of 24,000 genes have been

developed so far. Next we are (2) genotyping a *P. deltoides* unstructured population for association mapping. Oligonucleotides optimized for recovery of target coding and regulatory sequences are being used for sequence capture in 500 individuals of an association population. Captured fragments will be resequenced and polymorphisms genotyped for association analysis. (3) Upon completion of genotyping, we will identify significant SNP-trait associations with biomass growth and carbon partitioning to define genes and alleles that regulate trait variation. Alternative alleles detected in polymorphic sites will be tested individually and in a combined model for marker-trait association to identify the genes that regulate biomass growth and partitioning of carbon into lignocellulosics.

Poplars are the principal short rotation woody crop species for providing clean, renewable and sustainable fuels in North America, because of their fast, perennial growth habit and wide natural distribution in a broad range of environments. While poplars provide the benefits of an ideal bioenergy crop, with few exceptions, the genes that regulate productivity and biomass composition are largely unknown, despite their critical relevance for efficient conversion of biomass to biofuels. This gap is the main barrier for efficient molecular breeding and selection of superior poplar germplasm and, consequently, the extensive adoption of this woody crop as a renewable bioenergy source. Association genetics has become the primary approach for identification of genes that regulate complex traits in human genetics, agriculture and forestry because this strategy captures information on a broad range of alleles that control phenotypic variation, at high-resolution. Poplars are particularly suited to unveil the molecular basis of biomass productivity and composition using association genetics because of minimal domestication, large open-pollinated native populations with limited genetic structure, and high levels of genetic and phenotypic variation. However, with few exceptions, association genetic studies in plants have been hampered by limited gene and polymorphism coverage, because of the limited knowledge of the genetic variants and the low multiplexing capacity of genotyping platforms available to plant species. As a consequence, for most traits analyzed to date only a limited fraction of the genetic diversity impacting phenotypic variation has been uncovered.

189

Comparative Gene Expression of the *Caldicellulosiruptor* Genus using RNAseq

Loren J. Hauser[1]* (hauserlj@ornl.gov), Sara Blumer-Schuette,[2] Ira Kataeva,[3] Sung-Jae Yang,[3] Farris Poole,[3] Daniel Quest,[1] Inci Ozdemir,[2] Andrew Frock,[2] Erika Lindquist,[4] Tanya Woyke,[4] Bob Cottingham,[1] **Michael W.W. Adams,[3]** and **Robert M. Kelly[2]**

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]North Carolina State University, Raleigh; [3]University of Georgia, Athens; and [4]Joint Genome Institute, Walnut Creek, Calif.

Project Goals: This project has two goals. The first is to compare the gene expression profiles, using RNAseq, of a series of related high growth temperature bacteria from the genus *Caldicellulosiruptor* grown using both simple and complex carbon sources. The second is to develop a set of analysis tools to process large amounts of RNAseq data.

All known members of the *Caldicellulosiruptor* genus grow optimally between 65°C to 80°C and can anaerobically degrade plant biomass using various and complementary strategies. They are prime candidates for use in an industrial consolidated bioprocessing facility to produce second generation biofuels from complex plant material such as switchgrass. In collaboration with the Department of Energy Joint Genome Institute (JGI) we have recently completed sequencing and annotating the genomes of eight members of this genus. In addition, we have generated RNAseq data from four members grown on a variety of carbon sources including, glucose, maltose, cellobiose, starch, crystalline cellulose (Avicel), and dilute acid pre-treated switchgrass. Two of the primary advantages of RNAseq are its dynamic range and sensitivity. Greater than 98.5% of all protein coding genes had some detectable expression in all growth states and varied in expression level up to 10^6 fold. The expression levels of some genes, when grown on different carbon sources, varied by over 10^3 fold. As expected, the genes encoding ABC sugar transporters, cellulases and other glycosyl hydrolases were amongst the genes with the greatest changes in expression levels when grown on sugars versus complex carbon sources such as switchgrass. However, there were a number of other genes, such as members of a CRISPR cluster and some genes involved in fatty acid metabolism, that had unexpected changes in expression when grown on different carbon sources. We are developing an analysis pipeline to process and visualize the data and will also compare them with the results from DNA microarray analyses. RNAseq analyses will also include identifying the 5' end of transcription units, defining operons, identifying co-regulated genes and operons, and predicting transcription factor binding sites. Preliminary analysis has identified putative promoters embedded in genes, which allows the definition of unconventional operons and regulons. A thorough analysis will undoubtedly reveal additional unique biological phenomenon.

Biological Systems Interactions

190

PNNL Foundational Scientific Focus Area—Biological Systems Interactions

Jim Fredrickson[1]* (jim.fredrickson@pnl.gov) and **Margie Romine**[1]
Co-Principal Investigators: Gordon Anderson,[1] Scott Baker,[1] Alex Beliaev,[1] Bill Cannon,[1] Mary Lipton,[1] Jon Magnuson,[1] Thomas Squier,[1] and H. Steven Wiley[1]
Laboratory Research Manager: Harvey Bolton Jr.[1]
External Collaborators: Don Bryant,[2] Frank Collart,[3] William Inskeep,[4] Francois Lutzoni,[5] Andrei Osterman,[6] Margrethe Serres,[7] and David Ward[4]
PNNL Contributors: Alice Dohnalkova,[1] David Kennedy,[1] Bryan Linggi,[1] and Steve Lindemann[1]

[1]Pacific Northwest National Laboratory, Richland, Wash.; [2]Pennsylvania State University; [3]Argonne National Laboratory; [4]Montana State University; [5]Duke University; [6]Burnham Institute for Medical Research; and [7]Marine Biological Laboratory

Project Goals: The main scientific objectives of the PNNL FSFA include: development of a mechanistic understanding of interactions among key members of microbial autotroph-heterotroph associations (AHA) using the tools of genomics and systems biology; understanding the collective energy, carbon, and nutrient processing in AHAs that contributes to their stability and efficient utilization of resources; probing interspecies co-adaptations and functional innovations that contribute to robustness and functional efficiency and exploring the types of microbe-microbe and microbe-environment interactions that control genome evolution; and understanding cellular strategies that permit a system of interacting organisms to control the excess generation of reactive oxygen species to promote adaptive responses that enhance their survival.

The PNNL Genomic Science Foundational Scientific Focus Area (FSFA) is addressing critical scientific issues on microbial interactions, investigating how microorganisms interact to carry out, in a coordinated manner, complex biogeochemical processes such as the capture and transfer of light and chemical energy. The primary research emphasis is on associations between autotrophic and heterotrophic microorganisms with the additional objective of obtaining a predictive understanding of how interactions impart stability and resistance to stress, environmental fitness, and functional efficiency. The main scientific objectives of the FSFA include: development of a mechanistic understanding of interactions among key members of microbial autotroph-heterotroph associations (AHA) using the tools of genomics and systems biology; understanding the collective energy, carbon, and nutrient processing in AHAs that contributes to their stability and efficient utilization of resources; probing interspecies co-adaptations and functional innovations that contribute to robustness and functional efficiency and exploring the types of microbe-microbe and microbe-environment interactions that control genome evolution; and understanding cellular strategies that permit a system of interacting organisms to control the excess generation of reactive oxygen species to promote adaptive responses that enhance their survival (see PNNL FSFA posters for additional detail). Biological systems under investigation range from co-cultures consisting of well-characterized model organisms, to highly evolved lichen-forming microeukaryotes, to microbial mats associated with geothermal and hypersaline environments.

Microbial mats are highly organized, metabolically interactive, self-sustaining communities that develop in extreme environments. Hot Lake is a hypersaline, epsomitic ($MgSO_4$), shallow meromictic lake near Oroville, Washington that contains a benthic microbial mat (Figure 1A). Because of the high alkalinity and divalent cation concentrations of saline alkaline lakes, carbonate minerals can precipitate in these environments, often in association with microbial mats. These lakes therefore represent natural models for investigating carbon cycling and the microbial processes that may accelerate carbonate mineralization. The measured pH of Hot Lake varies from 8.5-9.2 and the Mg and SO_4^- concentrations can reach 1.7 and 2.2 molar, respectively, depending on season and depth. The top layer of the submerged mat (Figure 1B), immediately under the large crystals and on top of the green cyanobacterial layer contains magnesite ($MgCO_3$) crystallites that are encased in an amorphous layer of similar composition (i.e., Mg, C, and some Ca) (Figure 2). We hypothesize that organic polymers (e.g., extracellular polymeric substances) produced by the mat may be coating the particles, impeding crystallization at the surface.

Exploratory investigations of microbial diversity were conducted using pyrosequencing and revealed many taxonomic units consistent with phototrophy, chemoheterotrophy, sulfur cycling, and halotolerance. The most abundant phototrophs, based on nearest-neighbor phylogeny, included the filamentous cyanobacterium *Tychonema* and the phototrophic sulfur bacteria, *Halochromatium*. *Thiohalocapsula*, a purple sulfur bacterium, was also present as were purple non-sulfur bacteria affiliated with *Roseovarius* and *Rhodobaca*. Pyrosequencing also detectedmany phyla representative of halotolerant, sulfur-respiring bacteria including *Halothiobacillus* and sulfate-reducing bacteria including *Desulfonatronum* and *Desulfotignum*. Laboratory microcosms have been established from mat subsamples diluted into lake water to provide first-generation model systems where community structure and function responses to environmental variations can be measured. Future research plans include flow cytometry and sectioning in combination with metagenomics and metatranscriptomics to further characterize the functions and interactions among members of this mat community as a function of vertical position. In addition to systems biology investigations, we will also be investigating geomicrobial processes contributing to carbonate mineral formation.

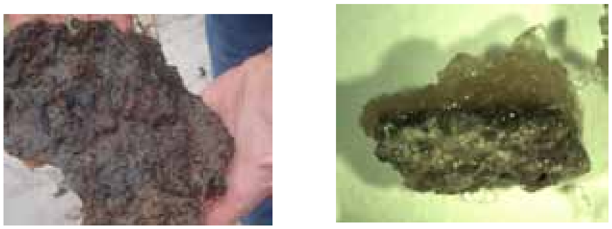


Figure 1. Hot Lake benthic microbial mat (top, mat underside) and vertical section illustrating layering.

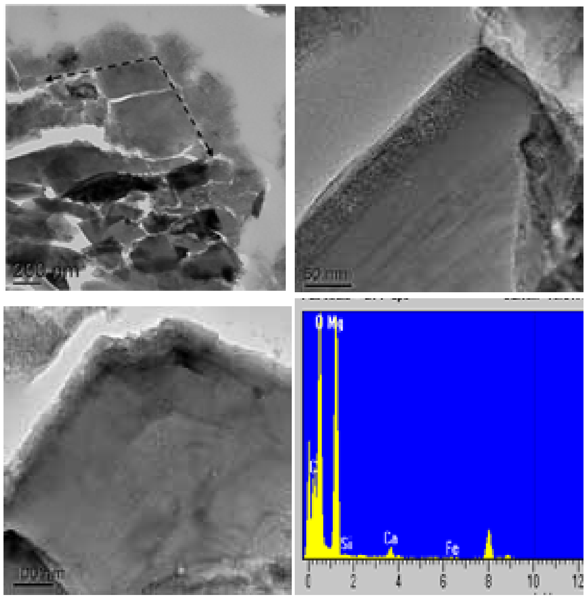


Figure 2. Crystallites from upper layer of the Hot Lake mat sampled just above the green cyanobacterial layer. The crystallites are magnesite ($MgCO_3$) as determined by selected area electron diffraction (SAED) with amorphous coatings of similar elemental composition as determined by energy dispersive X-ray spectroscopy (EDS).

This research was supported by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Genomic Science Program (GSP). This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

191

Mobile Gene Pools and the Functions They Encode Give Insight into How Genome Evolution is Shaped by Interactions Between Microbes and Their Environment

Margrethe Serres[1]* (mserres@mbl.edu), Sebastian Jaramillo Riveri,[2] and **Margie Romine**[2]

[1]Marine Biological Laboratory, Woods Hole, Mass.; and [2]Pacific Northwest National Laboratory, Richland, Wash.

Project goals: Differences in a microbe's ability to survive under various environmental conditions has been linked to the presence of genes conferring an advantage to its host. Some of these fitness genes have been obtained from other organisms residing in the same environment. Integron integrases represent one such mode of lateral gene transfer. The IntI integron integrase inserts foreign DNA through site-specific recombination between an AttI site in the integron and an AttC site present on a mobile DNA segment. The integron also encodes a promoter that facilitates the expression of the inserted gene(s). Detection of integron associated gene cassettes and the study of their encoded functions may give insight into how organisms enhance their functional efficiency and adapt to environmental factors. Analyzing the evolutionary history of the exchanged elements may also give insight into co-occurring organisms and their ability to affect genome evolution. The initial study is done on species belonging the genus *Shewanella* while future work will be directed at environmental samples of mats from Yellowstone National Park hot springs and meromictic saline lakes.

Current results

We are studying integron associated gene cassettes in the genus *Shewanella*. IntI integrases have been reported to be present in strains belonging to this genus. Also, as a group the *Shewanellas* have adapted to many environments that vary in their nutrient sources, salinity, temperature, and pressure. The genome sequences and ortholog gene sets are available for 21 closely and distantly related *Shewanella* strains.

One or two copies of IntI integron integrases were detected in 12 of the 21 *Shewanella* genomes. The IntI genome neighborhoods were searched for AttI and AttC integration sites. Sequence patterns of the identified *Shewanella* Att sites were then used for Blastn analyses of the *Shewanella* genomes. We found recombination sites with adjacent gene cassettes in all 21 genomes. A total of 1137 genes were identified as belonging to the integron associated gene cassettes. The number of integron cassette genes per genome ranged from 96 (*S. woodyii*) to 29 (*S. halifaxensis*). Genomes that encode two IntI genes had an overall higher number of integron integrase cassette genes. Approximately half of the integron gene cassettes consisted of one or two genes, with the remaining containing 3 to 14 consecutive genes. The largest gene cassettes had 12-14 genes, and these were found in the genomes of *S. amazonensis*, *S. baltica* OS195, *S. baltica* OS185, and *S. frigidimarina*.

The protein sequences encoded by the 1137 genes were analyzed for their functions. We found that 65% or the proteins were annotated as (conserved) hypothetical proteins. Others had functional descriptions with the most abundant being GCN5-related acetyl transferase, glutathione-dependent formaldehyde-activating enzyme, cytoplasmic adenylate cyclase, and glyoxylase family protein. To improve the annotation of the integron cassette encoded proteins we searched for conserved regions, or protein domains, with similarity to HMM models of proteins in the Pfam and TIGRFAM databases. Families of proteins with common domain(s) are known to encode similar functions. A blast analysis using trusted cutoffs detected similarity to proteins in the two databases for 879 of the integron cassette proteins. We detected 406 distinct Pfam and 355 distinct TIGRFAM domains, with the most prevalent domains encoding the most abundant protein functions listed above. The domain information is currently used to improve the annotation of the integron associated cassette proteins. Cellular roles associated with the TIGRFAMs indicate that the integron gene cassettes are enriched in proteins related to protein synthesis, cellular processes, energy metabolism, and regulation.

A blast analysis was done against the nr database to get insight into the evolutionary history of the laterally transferred gene pool. We did not include similarity to other *Shewanella* proteins as some of the integron cassette proteins had orthologs in the other *Shewanella* genomes.

The protein-pair with the lowest e-value was extracted for each of the integron cassette proteins. The most abundant sequence matches were to *Colwellia psyrerythraea* (40), *Pseudoalteromonas tunica* (38), *Alteromonas macleodii* (19), *Moritella* (18), and *Idiomarina loihiensis* (17). These microbes belong to the *Alteromonodales*, along with *Shewanella*. When counting sequence matches at the genus level, *Vibrio* (141) was significantly higher than the rest; *Pseudoalteromonas* (52), *Colwellia* (40), and *Pseudomonas* (35), likely reflecting the abundance of sequenced *Vibrio* genomes. This pattern of high similarity to *Vibrio* proteins did not differ when comparing different phylogenetic groups within *Shewanella* or when comparing proteins with or without *Shewanella* orthologs. Analyses of specific protein functions and of selected integron associated cassettes will be presented.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

192

Metagenome Analysis of High-Temperature Chemotrophic Microbial Communities Provides a Foundation for Dissecting Microbial Community Structure and Function

W. Inskeep[1]* (binskeep@montana.edu), M. Kozubal,[1] J. Beam,[1] Z. Jay,[1] R. Jennings,[1] H. Bernstein,[2] R. Carlson,[2] D. Rusch,[3] S. Tringe,[4] M. Romine,[5] R. Brown,[5] M. Lipton,[5] J. Moran,[5] H. Kreuzer,[5] C. Ehrhardt,[5] and J. Fredrickson[5]

[1]Department of Land Resources and Environmental Sciences and Thermal Biology Institute, Montana State University, Bozeman; [2]Department of Chemical and Biological Engineering, Montana State University, Bozeman; [3]J. Craig Venter Institute, Rockville, Md.; [4]Department of Energy-Joint Genome Institute, Walnut Creek, Calif.; [5]Department of Energy-Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: Use systems biology approaches to dissect community-level structure and function in geothermal microbial communities of YNP.

Microbial communities are a collection of interacting populations, each comprised of numerous individuals. However, a significant fraction of our knowledge base in microbiology originates from organisms grown and studied in pure culture, in the absence of other members of the community who may compete for resources or provide necessary cofactors and or substrates. Moreover, many of the organisms studied in pure culture have not represented the numerically dominant members of microbial communities found in situ. The advent of molecular tools (and -omics technologies) has provided opportunities for assessing the predominant and relevant indigenous organisms, as well as their likely function within a connected network of different populations (i.e., community). High-temperature microbial communities are often considerably less diverse than mesophilic environments and constrained by dominant geochemical attributes such as pH, dissolved oxygen, Fe, sulfide, and or trace elements including arsenic and mercury. Consequently, the goal of our work is to utilize high-temperature geothermal environments including acidic Fe-oxidizing communities as model systems for understanding microbial interactions among community members (Figure 1).

Recent metagenomic sequencing of high-temperature, acidic Fe-mats of Norris Geyser Basin, Yellowstone National Park (YNP) reveal communities dominated by novel archaea, members of the deeply-rooted bacterial Order Aquificales, and less-dominant Bacillales and Clostridiales. Phylogenetic and functional analysis of metagenome sequence is providing an excellent foundation for hypothesizing the role of individual populations in a network of interacting community members, and for testing specific hypotheses regarding the importance of biochemical pathways responsible for material and energy cycling. For example, we are using metagenome sequence in combination with information available from reference strains to identify protein-coding sequence of importance in the oxidation and or reduction of Fe, S, O, and As, as well as central C metabolism (including fixation of CO_2). Genes coding for proteins with hypothetical or putative roles in electron transfer, and C-cycling are being investigated using quantitative-reverse transcriptase-PCR (Q-RT-PCR) to evaluate mRNA levels in both pure-culture and mixed communities. Future transcriptomic and proteomic analyses will be coupled with detailed studies focused on the position of different organisms (spatial context) during Fe-mat development, as well as the role of O_2 flux across Fe-oxidizing boundary layers. Depositional studies have been conducted to correlate Fe-oxide deposition rates with O_2 flux rates measured using O_2-microelectrodes. Microelectrode measurements at the Fe-oxide-aqueous interface suggest significant O_2 consumption driven by the oxidation of Fe(II), but also show that sub-oxic conditions are common below the mat surface.

Metagenome sequence is being used to build consensus genome sequence of 5-6 dominant community members and will serve as a foundation for future sample analysis, laboratory data integration, and modeling efforts. Metabolic models for individual populations are constructed using elementary flux mode analysis, and these sub-components are then combined in a community model to explore possible ramifications of substrate interaction patterns on microbial community structure and function. Pathway specific processes are also being elucidated using isotope measurements focused on ^{13}C and ^{34}S of different chemical fractions with the goal of coupling this information to population-specific energy and nutrient cycling. Transcriptomic and proteomic results will be used to assess and confirm the importance of specific pathways and processes, and in conjunction with complementary datsets on C-isotopes and metabolomics, will allow refinements to individual and or community network models. Application of genomic, proteomic, and metabolic information to dissect microbial community structure and function is tractable within high-temperature geothermal systems, in part due to the relative simplicity of the communities and the stability of several key geochemical variables (i.e. pH, Fe, O_2).

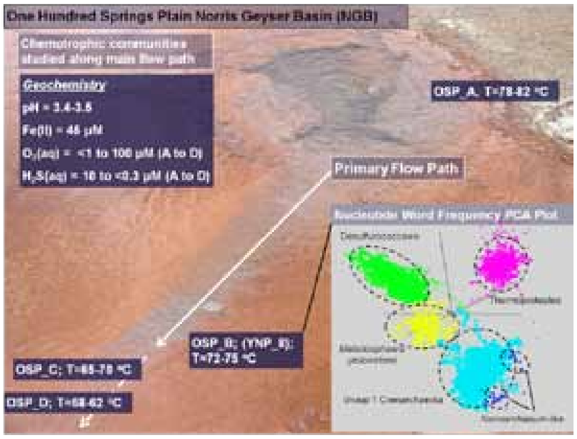

One Hundred Springs Plain Norris Geyser Basin (NGB)
Chemotrophic communities studied along main flow path
Geochemistry
pH = 3.4-3.6
Fe(II) = 46 µM
O2(aq) = <1 to 100 µM (A to D)
H2S(aq) = 10 to <0.3 µM (A to D)
OSP_A. T=78-82 °C
Primary Flow Path
Nucleotide Word Frequency PCA Plot
Desulfurococcales
Thermoproteales
Metallosphaera yellowstonensis
OSP_B; (YNP_B): T=72-75 °C
OSP_C; T=65-70 °C
OSP_D; T=58-62 °C

Figure 1. Site photograph of Fe-oxide depositing geothermal spring (One Hundred Spring Plain, OSP) located in Norris Geyser Basin (YNP). Reduced geothermal waters emerge with little to no detectable dissolved oxygen and significant levels of ferrous Fe (~50 µM). The oxidation of Fe(II) to produce amorphous Fe(III)-hydroxide is catalyzed by microbial populations such as *Metallosphaera yellowstonensis*. Inset at right indicates predominant microbial populations present at OSP_B (~72-75 C) identified using nucleotide word frequency plots of assembled metagenome sequence.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

193

Energy Carbon and Nutrient Partitioning in Lab-Based Phototroph-Heterotroph Co-Cultures

Alex S. Beliaev[1]* (alex.beliaev@pnl.gov), Grigoriy E. Pinchuk,[1] Oleg V. Geydebrekht,[2] Jennifer L. Reed,[2] Donald A. Bryant,[3] Allan E. Konopka,[1] Thomas Metz,[1] Sergey Stolyar,[1] and **Jim K. Fredrickson**[1]

[1]Pacific Northwest National Laboratory, Richland, Wash.; [2]University of Wisconsin-Madison; and [3]Pennsylvania State University, University Park

Project Goals: The overarching goal of the PNNL Foundational SFA is to understand the collective energy, carbon, and nutrient processing in phototrophic microbial communities that contributes to their stability and efficient utilization of resources.

Experimental systems for hypothesis testing include a well-defined engineered co-culture (e.g., *Synechococcus* 7002 – *Shewanella putrefaciens* W3-18-1), individual organisms (e.g., *Synechococcus* spp., *Thermosynechococcus* spp., *Roseiflexus* spp.), and consortia derived from or representative of naturally-occurring microbial communities (e.g., *Thermosynechococcus-Roseiflexus*). This researchis organized around three primary objectives which, in addition to science-driven tasks, encompass technical milestones thus enabling the transition from constructed opportunistic co-cultures to consortia derived from natural communities. The methods development efforts are focused primarily on analytical approaches for measuring biomass composition and mRNA/protein abundances in mixed cultures as well establishing techniques for identification of interactions between microorganisms. To identify secreted organic compounds that may serve as the primary carbon and energy sources for heterotrophic microbes growing in association with autotrophic species, we have tested the applicability of MS- and NMR-based detection techniques in conjunction with Chenomx metabolite library. While GC-MS had limited utility due to high concentration of interfering cationic salts, cNMR approach correctly identified and quantified dissolved organic acids, alcohols, and sugars in complex growth media.

To begin understanding the mechanisms governing the growth of photoautotroph-heterotroph associations, we are focusing on developing an integrated constraint-based flux balance model of the co-culture metabolism under environmentally relevant conditions. In initial studies, we used *Synechococcus* 7002-*S. putrefaciens* W3-18-1 co-culture to better understand the pathways of carbon and energy partitioning under different types of growth limitations relevant to natural associations which include limitations by carbon and light. Specifically, *Synechococcus* 7002 and *S. putrefaciens* W3-18-1 were successfully grown together in both batch and chemostat modes with lactate serving as the only source of carbon for both cultures. No mass-transfer of gases was applied suggesting that O_2 required to for complete lactate oxidation by *S. putrefaciens* W3-18-1 produced sufficient amounts of CO_2 to maintain growth of *Synechococcus* 7002. Our experiments also revealed that *Synechococcus* 7002 cannot grow in dark or in light using either lactate or acetate as the sole source of carbon while *S. putrefaciens* W3-18-1 cannot grow on lactate in the absence of O_2. Therefore, the co-culture utilizing light as the only source of energy and lactate as the sole source of carbon could only occur as a result of tight metabolic coupling between the phototrophic and heterotrophic organisms. *Synechococcus* 7002-*S. putrefaciens* W3-18-1 co-culture displayed balanced steady-state growth (μ=0.05 h^{-1}) under 1600 µmol/m^2/sec irradiance when dissolved oxygen tensions (dOT) were kept below 1%. However, under high dOT (160% of air saturation) and light intensity (2040 µmol/m^2/sec) the co-culture formed aggregates that primary consisted of filament-like *S. putrefaciens* W3-18-1 cells. While factors underlying these morphological changes are yet to be determined, we hypothesize that filamentous growth may play an important role in maintaining stable phototrophic biofilms which provide protection against reactive oxygen species. Notably, we have achieved sustainable growth of binary culture on inorganic source of carbon under light conditions. *S. putrefaciens* W3-18-1 was maintained in a chemostat mode for more than 15 generations using excreted photosynthate being the only carbon and energy source. Although initial experiments and methods development focus on opportunistic *Synechococcus* 7002-*S. putrefaciens* W3-18-1 co-culture, the naturally-occurring phototrophic mat communities provide an excellent opportunity to explore inter- and intra-guild metabolic interactions among specific populations associated with these communities. In a parallel effort, we initiated a study of a binary co-culture of *Thermocynechococcus* sp. NAK55 and *Roseiflexus castenholzii* DSM 13941; both organisms were isolated from phototrophic mat of Nakabusa hot spring (Japan). *Thermosynechococcus* NAK55 (kindly provided by Dr. S. Haruta) was successfully grown in liquid BG11 at 52°C without agitation under 20µE light. *R. castenholzii* DSM 13941 also grew on BG11 supplemented with 0.4% yeast extract in oxic and anoxic atmosphere in the dark and anoxic in the light at 52°C without agitation. Our current work is directed at revealing the nature of interactions as well as understanding the pathways of carbon, energy, and nutrient partitioning in both opportunistic and naturally occurring co-cultures.

An *in silico* model of metabolic coupling is also being constructed using *Synechococcus* 7002 and *S. putrefaciens* W3-18-1 constraint-based models. For the first assessment of *Synechococcus* 7002 growth *in silico* we used previously developed metabolic model for *Cyanothece* sp. ATCC 51142. It was estimated that the maximal biomass yields of *Synechococcus* 7002 was 0.024gAFDW/mmol CO_2 assuming CO_2 is the limiting substrate and the ratio of O_2 produced per CO_2 consumed at the maximal biomass yield was 1.05 O_2/CO_2. These parameters were used to approximate the behavior of *Synechococcus* sp. PCC 7002 in the binary culture. The metabolic model for *S. oneidensis* MR-1 (*i*SO783) was used to predict the behavior of *S. putrefaciens* W3-18-1 in co-culture, since 699 out of 783 genes in the MR-1 model have orthologs in W3-18-1. Most of the 29 reactions associated with the 74 missing orthologs from *S. putrefaciens* W3-18-1 were excluded from the model, except for three reactions that were essential. We estimated the growth associated ATP requirements for W3-18-1 assuming that it has 15% higher biomass yields than MR-1 as it was previously estimated. The model-predicted total biomass concentration in binary culture using 5 mM lactate as the sole source of carbon was in the range 0.3579 - 0.3863 gAFDW. Experimental assessment of chemostat lactate-limited binary culture growing in chemostat revealed that sum of total biomass and extracellular organic carbon was 0. 379 g/l, therefore confirming model prediction.

This research is part of the Foundational SFA at Pacific Northwest National Laboratory.

194

Decreased Protein Oxidative Damage Following Opportunistic Microbial Associations

Diana J. Bigelow,[1] Donald Bryant,[2] Baowei Chen,[1] **James K. Fredrickson,[1]** Na Fu,[1] Grigoriy E. Pinchuk,[1] Gaozhong Shen,[2] Thomas C. Squier[1]* (thomas.squier@pnl.gov), Sergey Stolyar,[1] and Yijia Xiong[1]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.; and [2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park

Project Goals: To identify stress resistance and adaptation mechanisms in microbial communities, we have the following goals:

- **Using detection assays against common oxidative modifications, use high-throughput screening approaches to explore how the efficient coupling of diverse metabolisms may prevent oxidative damage to biomolecules in model organisms *Shewanella* in culture with the cyanobacterium *Synechococcus*, and initiate measurements in natural systems involving lichins and co-cultures that reconstitute aspects of natural mat systems.**
- **Utilize advanced mass spectrometry capabilities available at PNNL to discover possible sites of oxidative and other post-translational modifications within microbial proteins from lysates isolated from axenic cultures of *Shewanella* and *Synechococcus*.**
- **Affinity isolate targeted oxidative modifications from natural populations as a function of the diel cycle, permitting the identification of likely sensors of oxidative stress in specific classes of bacteria.**

As part of PNNL's Scientific Focus Area "Biological Systems Interactions" this project is identifying regulatory proteins responsive to oxidative stress, which are hypothesized to play central roles in promoting stable associations between cyanobacteria and other microbes. Our strategy is to identify sequence-specific oxidative modifications in sensor proteins that could act to regulate metabolic fluxes and thereby control the excess generation of reactive oxygen species (ROS), which would promote adaptive responses that enhance cell survival. To identify fundamental adaptive strategies, initial efforts are focusing on laboratory model systems involving the cyanobacterium *Synechococcus* sp. PCC 7002 in culture with *Shewanella oneidensis* W3-18-1. We find that for cultures of the photoautotroph *Synechococcus* sp. PCC 7002 grown in the presence of the heterotroph *Shewanella* that there are large decreases in the overall levels of protein oxidation (Figure 1). Reductions in oxidative stress are apparent despite the substantially higher oxygen present in the co-culture (160% dissolved air saturation) in comparison to axenic cultures (44% dissolved air saturation) (both cultures are grown in caged bioreactors using white light intensities of 240 µmol photons m^{-2} s^{-1}) with shaking of 150 rpm. These results support the hypothesis that opportunistic interactions between heterotrophic (*Shewanella*) and photosynthetic (*Synechococcus*) microbes permit metabolic coupling to enhance energy efficiencies and community stability. Further, our results are consistent with prior indications that axenic isolates of the cyanobacterium *Synechococcus* sp. isolated from the microbial mat of Octopus Spring in YNP have a substantially enhanced sensitivity to light-induced oxidative stress in comparison to the natural mat community. This finding indicates an important role for metabolic coupling between community members in the mat that promote stress resistance.

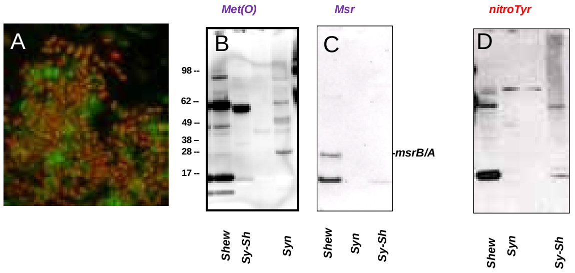


Figure 1. *Opportunistic Microbial Associations Diminish Protein Oxidation*. (Left) Image of co-culture showing *Synechococcus* (red) and *Shewanella* (green) (A). (Right) Immunoblots against methionine sulfoxide [Met(O)] (B), methionine sulfoxide reductase (Msr) (C), and nitrotyrosine (nitroTyr) (D) comparing extent of oxidative damage in co-culture (Sy-Sh) with that seen in axenic cultures of either *Synechococcus* (Syn) or *Shewanella* (Shew).

As hemes promote peroxidase and Fenton chemistries that can result in their oxidative modifications, we have examined the possibility that heme-containing proteins are oxidatively sensitive through a consideration of methionine sulfoxide formation. In the case of the major heme proteins, there is no significant methionine oxidation observed with the exception of one, fumarate reductase (Figure 2). The high level of methionine oxidation apparent for fumarate reductase, suggests possible linkages between oxidative stress conditions and alterations in metabolic flux that may arise due to oxidant-induced changes in fumarate metabolism. Given the positions of methionines within the fumarate reductase structure (SO0970; 1D4C.pdb), which occur within the active site and proximal to heme interfaces critical to efficient electron transfer, it is likely that methionine oxidation may functionally uncouple fumarate binding and its oxidation under aerobic conditions. As methionine oxidation is reversible through the actions of methionine sulfoxide reductases, these measurements suggest a means to maintain an ability to control the use of fumarate as an electron acceptor through its functional regulation in response to oxygen levels.

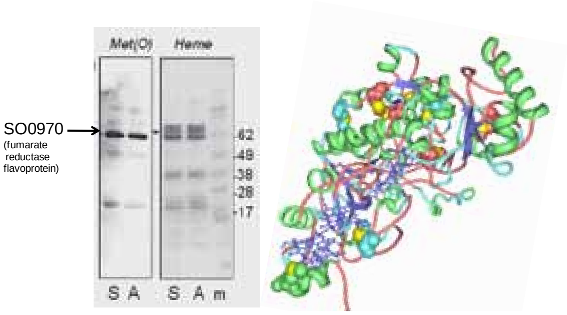


Figure 2: Targeted oxidation of methionines in furmarate reductase protein in Shewanella in the presence of suboxic (S) or aerobic (A) environmental conditions. (Left) Immunoblot against methionine sulfoxides [Met(O)] or heme stain (Heme) following SDS-PAGE protein separation. (Right) Structure of fumarate reductase (1D4C.pdb) highlighting positions of methionines (spacefilling) relative to hemes and FAD (purple stick representation). Colors of methionines correspond to backbone representation, where sulfurs are yellow.

Additional regulation through site-specific oxidative modifications involves the heme protein GlbN, which has been shown to protect *Synechococcus* sp. PCC 7002 against oxidative stress associated with growth on nitrate (Scott et al., 2010; Biochemistry 49, 7000). These growth conditions are suggested to mimic environmental conditions associated with denitrification, where cyanobacteria that express GlbN grow in association with soil organisms that produce nitric oxide as a result of anaerobic respiration using nitrate as an electron acceptor. We are examining the role of GlbN as an antioxidant protein against peroxynitrite, or possibly as a protein sensor that undergoes autocatalytic activation in response to oxidative stress, using mutant strains grown in the presence of high nitrate. Inspection of the GlbN structure (2KSC.pdb) indicates a close proximity between Tyr5 and Tyr22 and basic Lys side chains that we have previously demonstrated to result in the sensitive nitration of tyrosines in a range of proteins. Positions of these tyrosines within the tertiary structure suggest an ability to stabilize the fold of GlbN, which has the potential to create necessary binding pockets that enhance function against, for example, peroxynitrate. Isolation of the intact protein following growth on nitrate, in conjunction with *in vitro* measurements of function, offer a means to identify the role of GlbN and other sensor proteins that may function to control intracellular metabolisms that enhance community stability.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

195

Genomic Reconstruction of Vitamin Metabolism in Microbial Communities

Andrei Osterman[1]* (osterman@burnham.org), Dmitry Rodionov,[1] Leonardo Sorci,[1] Margaret Romine,[2] Jim K. Fredrickson,[2] Samantha Reed,[2] and Nadia Raffaelli[3]

[1]Burnham Institute for Medical Research, La Jolla, Calif.; [2]Pacific Northwest National Laboratory, Richland, Wash.; and [3]Università Politecnica delle Marche, Ancona, Italy

Project Goals: Genomics-based prediction and experimental validation of gene functions, pathways and networks in targeted groups of heterotrophic and phototrophic microbes for the fundamental understanding of microbial ecophysiology, evolution, adaptation and associations.

Metabolic cross-feeding is believed to play an important role in microbial communities. Whereas typical metabolic byproducts may provide a major flux of carbon and energy, vitamins (precursors of key cofactors) are required in relatively small amounts. Therefore, vitamin exchange may be rather widespread phenomena contributing to "opportunistic" relationships between species. This notion is consistent with our recent genomic survey, which confirmed that most environmental bacteria harbor both, de novo and salvage pathways for biogenesis of major vitamins (such as niacin, pantothenate, biotin, thiamin, riboflavin). Such species have a potential to benefit from as well as contribute to the vitamin pool in the environmental niche. At the same time, the observed mosaic distribution of *de novo* and salvage pathways leads to the presence in communities of the strict auxotrophs and strict prototrophs with respect to one or another vitamin. An ability to accurately reconstruct vitamin metabolic pathways and predict respective phenotypes over a rapidly growing collection of sequenced microbial genomes would impact our understanding of the metabolic cross-talk in microbial communities. To address this long-range goal we combine comparative genomics and predictive bioinformatic techniques with the experimental assessment of vitamin biochemical, transport and regulatory pathways in individual bacterial species and in model co-cultures. Using a subsystems-based approach captured in The SEED genomic platform, we were able to elucidate major vitamin biosynthetic and salvage pathways over a collection of >1,000 diverse bacterial genomes (seed-viewer.theseed.org). In addition to accurate projection of knowledge from model species to many others, this approach allows prediction of previously unknown genes and pathway variants. This is illustrated by the example of NAD metabolism, which is indispensable in nearly all analyzed species and reveals remarkable variations of associated pathways and regulatory mechanisms. Our recent analysis is focused on two model environmental bacteria, *Synechococcus* and *Shewanella*. A co-culture formed by the representatives of these groups, *Synechococcus sp. PCC 7002* and *Shewanella sp. W3-18-1* (under study at PNNL), provides a tractable model of microbial interactions between phototrophs and heterotrophs. Notably, both species appear to harbor divergent and unusual salvage pathways. Thus, *Synechococcus* sp. *PCC 7002* genome encodes an unprecedented (for Cyanobacteria) combination of PncA/PncB-mediated salvage of nicotinamide (Nm) with a possible NadR-driven utilization of ribosyl-Nm. At the same time, *Shewanella* sp. W3-18-1 appears to be one of the few species in this group with the unusual version of NadV-mediated Nm salvage pathway. A previously unknown gene encoding NMN deamidase (NadH) postulated for the second step of this pathway was cloned, and the respective purified recombinant enzyme was characterized. NadH orthologs are conserved in many diverse bacteria where they are likely involved in NMN recycling. The physiological role and the contribution of salvage genes and pathways to the overall NAD biogenesis are tested by a combination of genetic and metabolic profiling techniques. Remarkably, some (but not all) of the species in both groups, despite their taxonomic and metabolic distinctions, share a novel transcriptional regulator of NAD metabolism, NrtR. Regulons controlled by NrtR were analyzed as a part of our systematic genomic reconstruction of regulatory networks in *Shewanella* and Cyanobacteria and captured in the RegPrecise database (regprecise.lbl.gov/RegPrecise). Genomics-driven searches for additional regulatory mechanisms and presently unknown transporters involved in vitamin salvage and/or excretion are presently in progress.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

196

Global Metatranscriptomic Analyses of the Chlorophototrophic Microbial Mat Community of Mushroom Spring

Zhenfeng Liu,[1] Marcus Ludwig,[1] Christian G. Klatt,[3] David M. Ward,[3] and **Donald A. Bryant**[1,2]* (dab14@psu.edu)
Lead Principal Investigator: Jim K. Frederickson (Jim.Fredrickson@pnl.gov)

[1]Dept. of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park; [2]Dept. of Chemistry and Biochemistry, and [3]Dept. of Land Resources and Environmental Sciences, Montana State University, Bozeman

Project Goals: The overarching goals of our studies are to understand the composition of the microbial mat communities associated with alkaline siliceous hot springs, to understand the physiological capabilities of the individual populations of organisms, and to understand how these populations interact metabolically within the mat ecosystem.

The chlorophototrophic microbial mats of alkaline siliceous hot springs in Yellowstone National Park have served as models for studying the structure and function of microbial communities for decades. Culture-independent methods, including metagenomics and metatranscriptomics, have been used to define the composition and physiological potential of the organisms comprising these mats. These studies have showed that the mats of Mushroom and Octopus Spring are composed of 8 major populations, belonging to six kingdoms: *Cyanobacteria* (*Synechococcus* spp.), *Chloroflexi* (*Roseiflexus*, *Chloroflexus*, and *Anaerolinea*-like spp.), *Chlorobi*, *Acidobacteria*, *Firmicutes*, and *Bacteroidetes*. Three new chlorophototrophs (*Candidatus* Chloracidobacterium thermophilum, *Candidatus* Thermochlorobacter aerophilum, and a novel phototrophic member of the *Chloroflexi*) were discovered in these mats. The overarching goals of our studies are to understand the composition of this community, to understand the physiological capabilities of the individual populations of organisms, and to understand how these populations interact metabolically within the mat ecosystem. In this study the global transcriptome of the chlorophototrophic mat community of Mushroom Spring at 60-62°C was characterized.

Total RNA was extracted from chlorophototrophic mat samples collected at Mushroom Spring. Small RNAs (<300 nt) were removed and the remaining RNAs were reverse-transcribed to produce cDNAs. Four cDNA samples collected during light transition periods were initially sequenced by pyrosequencing (GX20 FLX) and SOLiD-3.5 technologies. The sequences produced by pyrosequencing were mainly used for rRNA analyses, while the SOLiD datasets were analyzed to discern gene regulation patterns. Sequences generated by pyrosequencing were aligned to rRNA databases using BLASTN. The composition of rRNA sequences was assessed from MEGAN analyses of the sequence alignments. The SOLiD datasets were aligned to assembled metagenome scaffolds (primary reference) and selected complete genomes of isolates from similar environments (secondary references) using the bwa algorithm and allowing 5 mismatches for the 50-bp sequences. Using artificially generated datasets, simulations of alignments allowing 0 to 5 mismatches showed that allowing 5 mismatches ensured a maximal number of correctly aligned sequences, a very low error rate, and a reasonable computation time. Test alignments using metagenome and complete genomes as references showed that complete genomes of biologically relevant isolates could serve as satisfactory references only when the sequence differences from the metagenome consensus were negligible. Gene expression patterns could be inferred by plotting normalized mRNA sequence counts for each gene as a function of time during a diel cycle. To minimize differences in sequencing coverage, organism abundance and cellular energy levels among samples, mRNA sequence counts for each gene were normalized to the total mRNA sequence counts of the organismal population to which the gene belonged. Finally, the resulting normalized data could be organized by the program ”Cluster” and visualized using “Java Treeview” to reveal groups of genes exhibiting transcription differences as a function of the diel cycle.

Members of four kingdoms (*Cyanobacteria, Chloroflexi, Acidobacteria*, and *Chlorobi*) accounted for ~84% of the rRNA sequences obtained by pyrosequencing. The remaining 16% of rRNA sequences showed similarity to many different rRNA sequence types. The most abundant sequences were similar to those of *Firmicutes* and *Bacteroidetes* (<2% each). Additional RNA samples were collected at 1-h intervals over a complete diel cycle. The amount of mRNA in cells varied significantly during the diel cycle and ranged from a low value of ~3% of the RNA sequences at night to ~12% in the late afternoon. For all chlorophototrophs in the mat, mRNA levels were highest during the midday when light intensity was highest. Changes in transcription occurred in all populations as a function of the diel cycle, and cluster analysis showed that transcripts could be assigned to two to four pattern classes for each organism during the course of the diel cycle. The transcription patterns (and the metabolism they represented) observed for oxygenic photoautotrophs were distinctly different from the patterns observed for aerobic anoxygenic photo(hetero) trophs (AAPs). Members of the *Cyanobacteria* were transcriptionally most active during the light period and maximally expressed their genes for components of the photosynthetic apparatus during the light period. As previously documented, expression of nitrogen fixation and fermentation-specific genes by cyanobacteria began in the late afternoon and continued until full sunlight again reached the mat in the morning. The global transcription patterns for members of the *Chloroflexi*, *Candidatus* Chloracidobacterium thermophilum and a *Chlorobiales* population were similar but differed from the patterns observed in the cyanobacteria. Transcript levels for genes encoding components of the photosynthetic apparatus peaked in *Roseiflexus* spp. populations were minimal during the daylight period but increased in the evening and were maximal in the early morning. Similarly, photosynthetic apparatus transcripts from *Candidatus C. thermophilum* and *T. aerophilum* peaked in the late afternoon and continued to be present at high levels throughout the night, but they were minimal during the daylight period. The transcription patterns suggested that factors other than light, most likely oxygen level, determine the periods of highest transcriptional activity for AAPs and specifically determine when the genes encoding components of the photosynthetic apparatus are expressed. The data provide an explanation for why organisms living in an environment characterized by very high light intensity have large antenna complexes such as chlorosomes: periods of high metabolic and transcriptional activity do not coincide with the periods of highest light intensity.

This microbial community is also being studied by microscopy, proteomics, and metabolomics to produce a complete model for the structure and metabolic interactions that occur in this complex but tractable model ecosystem.

This research was supported by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER’s Genomic Science Program (GSP). This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

197

Composition and Structure of Phototrophic Hot Spring Microbial Mat Communities: Natural Models for Systems Biology

David M. Ward[1]* (umbdw@montana.edu), Melanie C. Melendrez,[1] Eric D. Becraft,[1] Christian Klatt,[1] Jason Wood,[1] Donald A. Bryant,[2, 3] and Frederick M. Cohan[4]

[1]Land Resources and Environmental Sciences Department and [2]Dept. of Chemistry and Biochemistry, Montana State University, Bozeman; [3]Dept. of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park; [4]Department of Biology, Wesleyan Univ., Middletown, Conn.

Project Goals: Our overarching goal is to understand relationships between composition and function in a relatively simple natural phototrophic mat community inhabiting Yellowstone alkaline siliceous hot springs we have studied for many years in order to learn principles of how complex, multi-species biological systems work. Our specific goals are (i) identifying major populations using metagenomic assembly and recruitment with highly representative genomes, (ii) exploring species composition within these groups using rapidly evolving gene sequences or sets of gene sequences and theory-based evolutionary simulations to conduct single- and multi-locus population genetics analyses, (iii) examining the adaptations of different species through comparative genomics of representative isolates, (iv) observing patterns of community gene expression and protein synthesis through metatranscriptomic and metaproteomic analyses, (v) exploring ways to link important community functions, such as carbon sequestration.

Understanding how a complex biological system, such as a microbial community, works requires knowledge of its component microorganisms. It is essential that community composition be understood at the species level, as species are the fundamental taxonomic units that are uniquely distributed into distinct niches, make unique functional contributions and uniquely respond to the complex and dynamic natural environment. We have conducted long-term studies of microbial mats inhabiting alkaline siliceous hot springs, which are relatively simple, stable, high-biomass and very accessible systems that are protected within Yellowstone National Park. Our goal is to understand relationships between composition and function in this relatively simple natural community in order to learn principles of how complex biological systems work.

Initial impressions of community composition came from analyzing the genetic variation in 16S rRNA sequences. These studies revealed that the dominant native cyanobacteria (*Synechococcus* spp.), which play a major role in constructing the mat by performing oxygenic photosynthesis, are very unlike, and more diverse than, readily cultivated *Synechococcus* spp. isolates. Similarly, the dominant filamentous anoxygenic bacteria inhabiting the mat, which are thought to use infrared light energy to assimilate organic compounds excreted by *Synechococcus* spp., were found to be *Roseiflexus* spp., not the readily cultivated *Chloroflexus* spp.

The differential distribution of 16S rRNA variants along the effluent flow path (temperature gradient) and vertically in the upper portion of the mat photic zone (light gradient) suggested the presence of multiple, ecologically distinct *Synechococcus* populations. These populations might, like the Galapagos finches, have resulted from adaptive evolutionary radiation to fill distinct niches. However, analyses with more rapidly evolving genes, which offer higher molecular resolution, in combination with computer simulations of the evolution of ecological species, revealed that 16S rRNA sequence variation is unable to resolve all *Synechococcus* spp. in the mat. These approaches predict the existence of several tens of ecologically distinct *Synechococcus* spp., many of which exhibit unique spatial distributions, unique patterns of gene expression during the diel cycle and respond uniquely after imposing environmental change, as expected of ecological species. Cultivation-independent multi-locus sequence analyses, conducted by using bacterial artificial chromosome cloning to retrieve large genome segments of community members (*i.e.*, about 100 genes each), have demonstrated that genetic recombination among *Synechococcus* spp. has occurred, but has not been frequent enough to prevent the existence and detection of ecological species.

The relatively low genetic diversity of these mat communities has enabled the use of metagenomic technologies to describe its major types of inhabitants comprehensively. In particular, the low diversity permits assembly of paired-end sequences from individual metagenomic clones (2,000-12,000 nucleotides in length) into much longer contiguous genomic assemblies up to 1,600,000 nucleotides long), based on the similarity of overlapping regions of closely related sequences from the individual clones. These longer sequences are referred to as “metagenomic scaffolds”, where “meta-“ indicates that the scaffolds are comprised of sequences from the genomes of phylogenetically similar organisms, not a single genome. Scaffolds that have similar phylogenetic characteristics can be grouped together using k-means cluster analysis, which also separates scaffolds comprised of phylogenetically distinct sequences. Once annotated, these clustered scaffolds provide a means of linking genes that indicate phylogenetic affiliation with genes that give insight into the functional potential of the organisms represented by each cluster. Cluster analyses have revealed that 8 major populations predominate in the upper photic region of these mats. Clusters representing *Synechococcus* spp., *Roseiflexus* spp. and *Chloroflexus* spp. and a novel phototrophic acidobacterium (*Candidatus* Chloracidobacterium thermophilum), which we had discovered through analyses of functional gene sequences (photosystem reaction center genes) and later cultivated, were expected. A fifth cluster, confirmed the presence of a novel *Chlorobiales* population, which had been suspected based on the independent detection of 16S rRNA and reaction center sequences typical of members of the *Chlorobiales*. Three clusters represented predominant populations which had heretofore evaded detection, a remarkable finding for a system that has been so thoroughly investigated by numerous microbial ecologists for so long. One of these populations contains genes that indicate that it is a novel phototrophic member of Kingdom *Chloroflexi*. Its distant relationship to *Chloroflexus* spp. and *Roseiflexus* spp, suggests that phototrophy may have been an ancient phenotype in this kingdom. The 6 phototrophic clusters and their different gene contents suggest that diverse phototrophic guilds specialized to use different light wavelengths and autotrophic pathways, participate to coordinate efficient light capture and carbon sequestration in these mats. Two clusters, representing as yet unknown members of *Firmicutes*, and *Bacteroidetes*, do not contain genes indicating a potential to conduct phototrophy, and are likely contributed by heterotrophic members of the mat community, which are unrelated to heterotrophic bacteria previously cultivated from these mats.

Metagenomic assembly clusters appear to have coarse taxonomic resolution, lumping many individual phylogenetically related species. For instance, a single cluster is comprised of the many *Synechococcus* species described above. There is evidence that other clusters also contain multiple species, which, like *Synechococcus* spp., are likely adapted to distinct niches. We are using high-throughput sequencing approaches (specifically, Ti454 bar code analyses) to investigate species diversity of all 6 mat phototrophic populations simultaneously, based on variation in their photosystem reaction center genes. These analyses are being conducted on a large number and variety of samples, thus enabling simultaneous deep sampling of genetic diversity (thousands of sequences per sample), prediction of ecological species therefrom and validation of their unique ecological character. We are also in the process of obtaining deeper-coverage metagenomes for the entire mat community that will enable us to better understand the predominant populations involved in mat decomposition and carbon recycling.

This thorough understanding of community composition provides a solid basis for metatranscriptomic and metaproteomic analyses of gene expression and protein synthesis in the mats. The data from these systems-based approaches will also help us begin to understand and model how these species interact in space and time to coordinate community functions.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

198

Systems Biology of Lichen Systems: Pure Cultures of *Cladonia grayi* and Lichen-Dominated Biological Soil Crusts

PIs: **Daniele Armaleo,**[1] **Francois Lutzoni**,[1] **Frank Collart**,[2] **Scott Baker**[3]* (scott.baker@pnl.gov), and **Jon Magnuson**[3]* (jon.magnuson@pnl.gov)
Co-authors: Olafur Andresson,[4] Deanna Auberry,[3] Dave Culley,[3] Fred Dietrich,[1] Igor Grigoriev,[5] Brendan Hodkinson,[1] Steven Karpowicz,[6] Alan Kuo,[5] Peter Larsen,[2] Francis Martin,[7] Tami McDonald,[1] Sabeeha Merchant,[6] Emmanuelle Morin,[7] Olaf Mueller,[1] Ellen Panisko,[3] Maria Virginia Sanchez,[8] Ian Small,[9] and Basil Britto Xavier[4]

[1]Duke University, Durham, N.C.; [2]Argonne National Laboratory, Argonne, Ill.; [3]Pacific Northwest National Laboratory, Richland Wash.; [4]University of Iceland; Reykjavik, Iceland; [5]Joint Genome Institute, Walnut Creek, Calif.; [6]University of California, Los Angeles; [7]Institut National de la Recherche Agronomique, Nancy, France; and [8]Universidad Nacional de Cuyo, Mendoza, Argentina; and [9]The University of Western Australia, Crawley

Lichens arguably represent the most successful symbiotic relationship in nature. A lichen is composed of a heterotrophic fungus with a phototrophic partner; a green alga or a cyanobacterium or sometimes both. This project is a collaboration of Duke University, Argonne National Laboratory (ANL) and Pacific Northwest National Laboratory (PNNL) to explore the systems biology of lichens at two extremes of natural complexity; the partners of one species of lichen, *Cladonia grayi*, and complex biological soil crust (BSC) communities dominated by lichens. Lichens are found in all environments, and wherever they are found they have an important role in the ecosystem. They are sensitive to environmental changes. Therefore, they may be valuable sentinels for assessing the effects of climate change. Though slow growing, they are important in carbon flux in many environments, e.g., in BSCs and tundra. From a fundamental scientific standpoint lichens represent the best and most diverse examples of co-evolved phototrophic-heterotrophic systems.

The principal goal of this project with respect to *Cladonia grayi* is to use systems biology and molecular biology tools to advance fundamental understanding of the lichen symbiosis. Both symbiotic partners of *C. grayi* have been successfully cultured in isolation by the Duke University team members. This has enabled them to sequence the genome of each partner, the fungus *C. grayi*, and the green alga *Asterochloris* sp. Together with JGI and many others the assembly and annotation of the genomes has begun. Furthermore, the ability to culture the partner organisms separately and reestablish the early stages of the symbiotic relationship provides an unprecedented opportunity to study the biology of lichens using modern systems biology tools.

The genomes of the symbiotic partners from the lichen *Cladonia grayi* were assembled from 454 and Solexa data and partially annotated. The data, at JGI (http://genome.jgi-psf.org/Clagr2/Clagr2.home.html), have not yet been released publicly. The 40 Mb genome of the fungal partner (*C. grayi*) includes ~12,000 gene models and the 56 Mb genome of the green algal partner *Asterochloris sp*. has ~9,000 gene models. Signatures of symbiosis are being sought by surveying expansion or contraction of gene families, genes of ancient origin under stabilizing selection, organelle genomes, and gene expression during early development. A peculiar class of ammonium transporters unique to lichens and land plants was identified, as well as a polyketide synthase gene cluster responsible for lichen-specific compounds. The prokaryotic communities associated with *C. grayi* have also been surveyed. Transcriptomic investigations of the various cultured states are underway at ANL to enable the generation of predictive models relevant to complex environmental systems under conditions of climate change. A task that preceded the transcriptomic work was the generation of a map of the JGI annotated enzyme activities for *Cladonia* and *Asterochloris* to the KEGG map01100 (complete metabolism network). ANL has generated a series of interactive maps for the individual organisms as well as combined map that illustrates annotated enzyme classification (EC) activities that are unique to the individual organisms. Proteomics studies of liquid and plate cultures of the individual symbiotic partners from *Cladonia grayi* have begun at PNNL. The initial emphasis has been on the development of methods for cell lysis and proteomic analysis of the two organisms and determining the minimum amount of biomass required for a proteomic sample. The latter point is important in examining the early stages of the symbiosis establishment in the Duke culturing system, where biomass quantities are limited. Analysis of the initial LC-MS proteomics data for the fungal partner is underway and will be presented.

The goals of this project with respect to the BSCs are: to examine the phylogeny of selected lichen-dominated BSC communities obtained from arid land ecosystems; and apply transcriptomics and/or proteomics techniques to investigate these BSCs under different physiological states, e.g., naturally desiccated and rehydrated. BSCs are a remarkable and extremely sensitive type of microbial community important for retention and conditioning of soil and providing an environment where arid land plants can germinate and prosper. They may be an important indicator of climate change and undoubtedly impact carbon flux in arid environments throughout the world. A site for collection of BSCs has been established on Bureau of Land Management land in south-central Washington state. The study site has BSCs of various stages of maturity in a shrub-steppe ecosystem on land with steep topography that has not been used for agriculture or burned over in recent time. Initial studies of the phylogeny of these communities have revealed the genus of two of the major lichens, *Lecanora* and *Caloplaca* spp., in the selected BSCs. Broader studies of the eukaryotic and prokaryotic community will be conducted in the near future.

This research is part of the Genome Sciences Program FSFA at Pacific Northwest National Laboratory.

199

Isotopic Studies of Biological Systems

James Fredrickson and **Helen Kreuzer* (**Helen.kreuzer@pnl.gov)

Pacific Northwest National Laboratory

Project Goals: This project has two parts. The goal of the first part is to use hydrogen stable isotope measurements to elucidate biological hydrogen production pathways. The goal of the second part is to use carbon stable isotope measurements to trace carbon flow through lithoautotrophic acid hot springs microbial communities in Yellowstone National Park.

1. Using H Isotopes to Elucidate Biological Hydrogen Production Pathways

Eric L. Hegg,[1]* H. Yang,[1] N. Ostrom,[1] H. Ghandi,[1] H. Kreuzer,[2] E. Hill,[2] and J. Moran[2]

[1]Michigan State University; and [2]Pacific Northwest National Laboratory

Biological H_2 production by hydrogenase (H_2ase) enzymes has enormous potential as an environmentally sustainable energy source. H_2ases, found throughout nature in many diverse organisms, are among the most efficient H_2-producing catalysts known. Although considerable progress has been made in elucidating the metabolic pathways involved in H_2 metabolism, one major impediment to improving our understanding of H_2 metabolism is our inability to adequately define the regulation of and the flux through key pathways involved in H_2 production. **The goal of this project is to develop stable isotopic approaches for improving understanding of biological H_2 production.**

We predicted that the isotope ratio of H_2 produced by various H_2ases would differ because of slight differences in their active sites and proton transfer pathways. We further predicted that we could measure this difference via isotope-ratio mass spectrometry, and that the H/D isotope ratios would allow us to address fundamental questions concerning biological H_2 production. To test this predictions, we purified five different H_2ases [three [FeFe]-H_2ases (*Clostridium pasteurianum, Shewanella oneidensis,* and *Chlamydomonas reinhardtii*) and two [NiFe]-H_2ases (*S. oneidensis* and *Desulfovibrio fructosovorans*)] and established conditions that allowed us to quantify the specific activity of the purified H_2ases, the amount of H_2, and its isotopic content. Using the enzymes and optimized protocols established above, we determined the isotope ratio of the H_2 produced by the purified Significantly, the data indicate that all five H_2ases produce H_2 with a unique isotopic signature, demonstrating that different H_2-producing enzymes have different fractionation factors reflected in the isotope ratio of the H_2.

Shewanella oneidensis MR-1 is a facultative anaerobe capable of transferring electrons to a variety of terminal acceptors including iron, manganese, and other metals. *S. oneidensis* encodes two H_2ases, the [FeFe]-H_2ase HydA and the [NiFe]-H_2ase HyaB. We measured the isotopic content of H_2 produced in the headspace above cultures that had either the [FeFe]-H_2ase or the [NiFe]-H_2ase deleted, and we found that the isotope ratios closely matched those predicted by the in vitro studies with the purified enzymes. We next monitored both concentration and the isotope ratio of the H_2 gas in the headspace of *S. oneidensis* cultures. We grew MR-1 in 69 mL sealed serum bottles containing 20 mL of the defined growth medium M1. We used 60 mM lactate as the electron donor, and oxygen (from air-saturated water) as the electron acceptor. H_2 gas production began at the point of electron acceptor limitation in the serum bottles and went through two distinct phases, initial and late production. Although mRNA from both H_2ases was present during each phase of H_2 production, isotope ratio data indicated that initial H_2 production, starting at about 20 hours, was by the HyaB (Ni-Fe) H_2ase. This H_2ase then consumed H_2 from about 50 to 100 hours after inoculation. Late phase H_2 production after 100 hours was driven by the HydA H_2ase.

2. Carbon Flow in Lithoautotrophic Acid Hot Springs Microbial Communities, Yellowstone National Park

Helen Kreuzer,[1] James Moran,[1] Christopher Ehrhardt,[1] and Bill Inskeep[2]

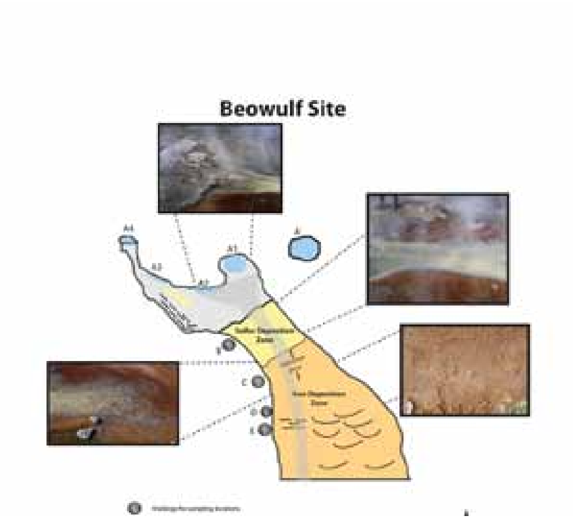
[1]Pacific Northwest National Laboratory; and [2]Montana State University

Beowulf Springs and Obsidian Pool in Yellowstone National Park are acidic hot springs. Beowulf is a sulfidic spring, while Obsidian Pool is not. The microbial communities in the hottest portions of these springs are thought to be lithoautotrophic, but carbon flow in these communities has not been characterized. Geochemical analysis shows that there are several potential sources of C present, including CO_2, methane, dissolved inorganic carbon, and dissolved organic carbon. Although CO_2 has been assumed to be the ultimate C source, any of these other C-containing materials could potentially be a C source, and the topography of Obsidian Pool suggests it might receive heterotrophic C input from the surrounding landscape. **The goal of this project is to trace C flow through these communities.**

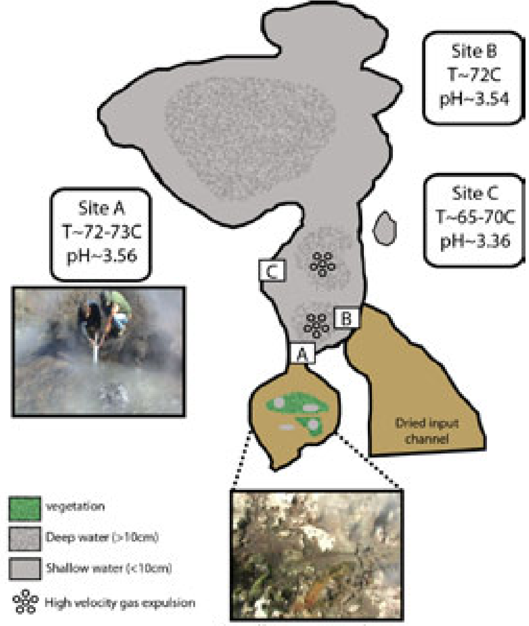
We hypothesize that the C sources will have different C isotope content, enabling an initial tracing of C into the community through isotope profiling. We collected samples from each of the sites (see figure), and are in the process of analyzing the C stable isotopic content of dissolved methane, outgassed CO_2, dissolved organic and inorganic carbon, fatty acid methyl esters and archaeal lipids from the in situ microbial communities, as well as potential organic substrates from surrounding environmental inputs. We will use lipid isotope data to associate potential C sources with phylogenetic groups of organisms and base the design of future stable isotope probing experiments on these results. In a collaborative effort (T. Woyke), samples from the same Obsidian Pool sites and one of the Beowulf sites are undergoing cell sorting at the JGI. If successful, we may be generating single-cell sequence data that can be coupled with the isotope data to help better link and resolve carbon flow in these communities.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

Beowulf Springs



Obsidian Pool



200
The Development of RNA-Sequencing Technologies for the Analysis of Complex Microbial Communities

Bryan E. Linggi,[1] Meng L. Markillie,[2] Chris S. Oehmen,[3] Ronald C. Taylor,[3] and **H. Steven Wiley**[1]* (steven.wiley@pnl.gov)

[1]Environmental Molecular Sciences Laboratory, [2]Biological Sciences Division, [3]Computational Biosciences, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: Develop a suite of technologies to enable the accurate measurement of molecular species (DNA, RNA, protein, metabolites) in microbial communities as well as building the computational infrastructure needed to analyze the resulting data.

Microbial communities are complex systems containing 10s to 100s of different organisms performing a variety of interrelated functions. To understand and develop ways to utilize these communities for human benefit, technologies are needed to probe their function and accurately measure their composition as well as the biochemical and molecular species that are present. To address this challenge, we are developing a suite of technologies to enable the accurate measurement of molecular species (DNA, RNA, protein, metabolites) in microbial communities as well as building the computational infrastructure needed to analyze the resulting data. As test cases, we are examining both 'synthetic' as well as natural microbial communities.

Metatranscriptomics

At the broadest scale, 'omics' experiments, such as genomics, transcriptomics, proteomics and metabolomics, can provide estimates of both the species composition of communities as well as the expression levels of key metabolic pathways. Ideally, the integration of multiple types of 'omics data could provide a comprehensive profile of community structure and function, but there are technical difficulties and tradeoffs with different technologies that complicated direct comparison of analytical results. For example, transcriptional profiling through sequencing-based technologies (RNA-Seq) requires removal of ribosomes to enable adequate sampling of complex microbial communities. In addition, RNA extraction efficiency can vary by organism, which could bias the estimated abundance of certain transcripts. To develop approaches to address these types of complexities, we have created an artificial community comprised of a mixture of 12 different species (12 bacteria, 2 archaea) in which the origin of the different transcripts could be identified though a ligated "barcode" sequence. Isolated RNA from each species was mixed in defined proportions and then sequenced together using the ABI SOLiD4 sequencing platform. We then tested different analytical and computational approaches to determine which combination could most accurately identify the source and abundance of each RNA fragment. We found that even direct matching of each read to its source genome generates a fraction of unmapped reads. To discriminate between sequences that don't match known genomes because of sequencing errors or inefficient mapping algorithm versus those that arise from an unknown organism, we generated a "gold standard" match for each sequence using ScalaBlast running on the EMSL Chinook supercomputer. While computationally demanding, ScalaBlast provides a much more reliable match relative to RNA-Seq alignment algorithms such as Bowtie, which sacrifice accuracy for speed. A number of other read matching algorithms were also tested to determine their relative accuracy. This allowed us to determine the optimal tools to use for metatranscriptome matching.

Another issue that must be resolved when using RNA-Seq technology for community profiling is the extent of ribosome removal that is necessary. Because of the complexity of microbial communities, extensive sampling is required to ensure unbiased coverage. Unfortunately, most of the RNA in microbes is ribosomal (rRNA), which is usually non-informative with respect to community function. If the ribosomes are not removed, estimates of transcript abundance between different organisms could be erroneous. Removal of the rRNA, however, could also potentially bias mRNA representation. Thus, we benchmarked several different approaches for rRNA removal to determine the extent to which different methods actually alter mRNA composition. Data from both the rRNA removal experiments and sequencing matching trials were used to design optimal strategies for transcriptional profiling of microbial communities.

Computational Infrastructure

The acquisition of large amounts of both proteomics and transcriptomics data from microbial communities presents difficulties for data query and analysis. However, effective integration of these types of data could provide a wealth of information on control of protein levels and post-transcriptional regulatory mechanisms. To address this, we recently completed a pilot project for development of a large-scale data warehouse / workspace for the analysis of extremely large data sets. The workspace was constructed as a terabyte-size parallel processing platform based on the Hadoop / MapReduce / HBase framework, implemented on one cluster with a distributed file system. Sample data were generated by performing comprehensive RNA-Seq and proteomics analysis of *Shewanella* grown under aerobic and oxygen-limited conditions in chemostats. The data consisted of over 29,000 different peptides observed in 4 different runs of the two samples. RNA-Seq runs provided transcript data in duplicate from each sample. Following data normalization, we matched RNA data that corresponded to each observed peptide in each sample and calculated a peptide/RNA ratio for each observed peptide. Confidence values and statistics were assigned to each peptide/RNA ratio across duplicates and across conditions. Preliminary analysis of the data suggests a strong correlation between the level of mRNA and most proteins, but that there are several membrane proteins that behave anomalously. We found that several technical issues must be addressed before peptide/RNA ratios can become a reliable measure of post-transcriptional regulation, such as correcting for the different dynamic ranges of RNA-Seq and proteomics measurements and developing appropriate error models for combined 'omic measurements. Nevertheless, we were successful in creating a robust computational infrastructure for rapidly integrating large amounts of genome-centric data and this should greatly facilitate its exploration and use in defining microbial regulatory processes.

This research is part of the Genomic Science Program FSFA at Pacific Northwest National Laboratory.

201

Information and Informatics Resource for Collaborative Research on Biological Systems Interactions

Gordon Anderson,[1]* William Cannon,[1]* Lee Ann McCue,[1]* Zoë Guillen,[1] Linda Angel,[1] Sebastian Jaramillo Riveri,[1] Margrethe Serres,[2] Margie Romine,[1] and **Jim Fredrickson**[1]

[1]Pacific Northwest National Laboratory, Richland, Wash.; and [2]Marine Biological Laboratory, Woods Hole, Mass.

Project Goals: As part of the Foundational Scientific Focus Area research program on Biological Systems Interactions at the Pacific Northwest National Laboratory, we have developed an information and informatics resource for collaborative research. The resource provides a computational infrastructure for information dissemination, analysis, discussion and data sharing. The resource currently combines three core technologies to enable collaborative research: semantic wiki technologies, an open source content management-based global file system, and pathway-genome databases and tools. The url for the Biological Systems Interactions research program is microbes.pnl.gov.

Wiki-enabled Collaborations

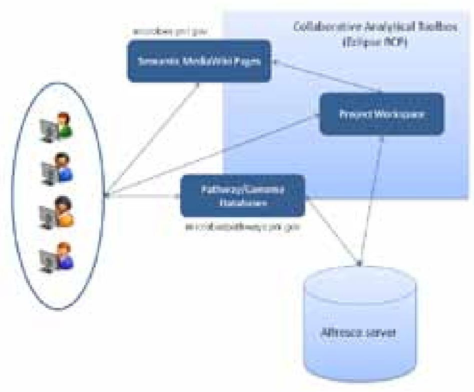
Research projects need the ability to collect and exchange both structured and ad hoc information collectively. We are using semantic wiki technology to allow researchers to create and edit interlinked wiki pages with ad hoc descriptive information about their experiments. A researcher can create a wiki page to provide the background information, research goals, hypotheses, and experimental outcomes of their project. The wiki also makes it easy to show the relationships between the projects in our research program by linking the project description wiki pages for related projects. We are also working to integrate the wiki with our content management-based global file system (CAT), so that there is a direct link between the ad hoc description of an experiment and its associated structured data stored in CAT.

Collaborative Analytical Toolbox (CAT)

CAT, developed by Pacific Northwest National Laboratory, is a client-server analytic framework for building and organizing a knowledge repository. The CAT environment provides a familiar, customizable interface that enables users to:

- build, organize and share their information;
- view/browse their information in any number of ways and via any number of arbitrary hierarchies;
- browse or search for information from a variety of sources using many different search tools;
- pull data back into their project space;
- integrate with existing tools to analyze their data;
- collaborate with other users by sharing data, templates, annotations, etc.

CAT was designed with flexible integration as a key requirement, so that it is easy for new or existing tools to be integrated with minimal development cost. Specifically, it is based on well-documented, well-supported, robust Eclipse Rich Client Platform (RCP), using the Alfresco content management system on the server side.



The RCP framework is a cross-platform architecture for building and deploying rich client applications. This framework is extendable through plug in applications to extend its functionally and customize it for the application domain. Our developments will extend CAT through the addition of new functionally and interfaces such as:

- Genome and ortholog interactive editor / curation tools
- Pathway and genome databases integration
- Enhanced features – links to KEGG/SEED/UniProt/IMG

Additionally, we are investigating integration of Alfresco content management with MediaWiki to support seamless integration of the information in CAT and the Wiki. This poster will present our current architecture as well as illustrate our vision for information assimilation designed to enable collaborative research across institutional boundaries.

Pathway and Genome Databases

Integration of the Collaborative Analytical Toolbox with pathway databases will greatly facilitate model building. Refinement of gene annotations is a continual process and takes place both prior to and after automated pathway prediction. However, the refinement of gene annotations and pathway annotations need to be coordinated, because refinement of gene annotations may result in the need to reevaluate the putative pathway annotations. The coordination between these activities is achieved within CAT. In addition, the initial pathway models must be evaluated with respect to predictions from other sources. Continuous evaluation of the pathway annotations is aided by the construction of a workflow that draws data from multiple sources such as MetaCyc, KEGG, IMG, and SEED, and allows for side-by-side comparisons of pathway models and membership. The revisions to the models are captured in BioCyc pathway-genome databases.

Currently available pathway databases include

- *Chloroflexus aggregans* DSM 9485
- *Chloroflexus* sp. Y-400-fl
- *Roseiflexus castenholzii* DSM 13941
- *Roseiflexus* sp. RS-1
- *Shewanella oneidensis* MR-1
- *Shewanella* sp. W3-18-1
- *Synechococcus* sp. JA-2-3B'a(2-13)
- *Synechococcus* sp JA-3-3Ab
- *Synechococcus* sp. PCC 7002
- *Synechocystis* sp. PCC 6803
- *Thermosynechococcus elongatus* BP-1

Support: This research is part of the GSP FSFA at Pacific Northwest National Laboratory

Plant-Microbe Interfaces

202

Plant-Microbe Interfaces

Mitchel J. Doktycz[1*] (doktyczmj@ornl.gov), Gerald A. Tuskan,[1] Christopher W. Schadt,[1] Gregory B. Hurst,[2] Edward Uberbacher,[1] Dale A. Pelletier,[1] Timothy J. Tschaplinski,[1] Francis Martin,[4] Rytas Vilgalys,[5] Amy Schaefer,[6] Caroline Harwood,[6] Jennifer Morrell-Falvey,[1] David J. Weston,[1] Scott T. Retterer,[1] Andrey Gorin,[3] Yunfeng Yang,[1] Robert Hettich,[2] Udaya C. Kalluri,[1] Xiaohan Yang,[1] Abhijit Karve,[1] Mircea Podar,[1] Steven D. Brown,[1] Robert Cottingham,[1] Tatiana Karpinets,[1] Chongle Pan,[3] Guru Kora,[3] and Susan Holladay[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [4]INRA, Nancy, Champenoux, France; [5]Biology Department, Duke University, Durham N.C.; and [6]Department of Microbiology, University of Washington, Seattle

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Rapid progress in biological and environmental sciences has been enabled by the availability of genome sequences and the tools and technologies involved in interpreting genome function. As our understanding of biological systems grows, it becomes increasingly clear that the functional expression of individual genomes is affected by an organism's environment and the community of organisms with which it associates. The beneficial association between plants and microbes exemplifies a complex, multi-organism system that is shaped by the participating organisms and the environmental forces acting upon it. These plant–microbe interactions can benefit plant health and biomass production by affecting nutrient uptake, influencing hormone signaling, effecting water and element cycling in the rhizosphere, or conferring resistance to pathogens. Studying the integral plant–microbe system in native, perennial plant environments, such as *Populus* and its associated microbial community, provides the greatest opportunity for discovering plant–microbial system functions relevant to DOE missions related to bioenergy and carbon-cycle research and understanding of ecosystem processes.

The functional attributes of *Populus* depend on the microbial communities with which it associates. Bacteria and fungi can be found within *Populus* tissues and closely associated with the roots in the rhizosphere. Understanding these communities, and the interfaces between organisms, is critical to realizing fundamental scientific knowledge that may enable increased plant productivity, ecosystem sustainability, disease resistance, drought tolerance, and ecosystem carbon budgets. This interface can also influence the processes, or mechanisms, by which adaptive traits arise from genetic variation and community function. Microbial rhizosphere structure, plant root bacterial and fungal colonization patterns, and the microbe–plant signaling pathways inherent in each type of association are all found within *Populus* and can be functionally translated hierarchically across scales into ecosystem patterns and processes.

Understanding the mechanisms by which plants and microbes interact represents a grand challenge facing biological and environmental science. How microbial selection and colonization occurs, what reciprocal benefits are bestowed upon the plant and microbe, and how these interactions ultimately affect, and are affected by, the environment are just some of the intrinsic scientific questions. The multiple spatial and temporal scales involved in these interfaces, the complexity of the component systems, and the need for better tools that use and build upon growing genomics resources to probe and interpret these combined systems represent some of the essential technical challenges. An overview of the research being carried out in the ORNL Plant Microbe Interfaces science focus area will be presented.

203

Plant-Microbe Interfaces: Proteomics Studies of Plant-Microbe Interfaces

Karuna Chourey[1] * (choureyk@ornl.gov), Xiaohan Yang,[2] Ting Li,[2] Chongle Pan,[1,3]Zhou Li,[4] Rachel M. Adams,[4] Robert L. Hettich,[1] Patricia K. Lankford,[2] Keiji G. Asano,[1] Andrey A. Gorin,[3] Udaya C. Kalluri,[2] Poornima Sukumar,[2] David J. Weston,[2] Sara M. Allen,[2] Dale A. Pelletier,[2] Timothy J. Tschaplinski,[2] **Gerald A. Tuskan**,[2] **Mitchel J. Doktycz**,[2] and **Gregory B. Hurst**[1]

[1]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]BioSciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [4]UT-ORNL Graduate School of Genome Science and Technology, Knoxville, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Characterization of proteomes of organisms and communities has the potential to contribute substantially to our understanding of plant-microbe interfaces, in particular the interactions among *Populus* and its microbiome. To realize this potential, a number of challenges posed by these complex systems must be overcome, requiring the development of tools and protocols that will extend the state of the art in proteomics. These challenges include (1) the complexity of microbial communities, and the lack of readily available metagenomes (2) partial protein extraction from plant and fungal tissues, often impeded by interfering small compounds, and protein degradation by endogenous enzymes, (3) collection of extracellular proteins that may play roles in signaling, (4) informatics associated with proteomics of eukaryotic organisms (plant, fungal) for which genomes reflect gene duplication events, alternate splicing, etc. A major future challenge for which we are currently designing approaches is the metaproteomics of *Populus* root microbial communities, including rhizosphere and endophyte components. To address these challenges, we have initiated proteomics studies in several critical areas for the Plant-Microbe Interfaces (PMI)-SFA:

-Analysis of small proteins encoded by small genes in bacterial and plant species To complement work pioneered by Xiaohan Yang to identify small-protein-encoding genes (see poster by Li et al.), we are implementing size fractionation to enrich small proteins from plant tissues in order to increase the sensitivity of LC-MS-MS analysis toward these analytes. Proteins identified by LC-MS-MS from several fractionation methods exhibited molecular mass distributions with medians significantly lower than those of unfractionated lysates. A number of the identified proteins were unique to small protein enrichment fractions. Among these proteins are several with annotations indicating that their functions are not yet characterized. Evidence for expression of these proteins supports improved annotation of the corresponding small genes, and provides candidates for further studies of biological function.

-Proteomics studies of roots from field-sampled mature *Populus* trees The ability to study the root proteome of plants is the first step towards unraveling the plant-microbe interactions in the rhizosphere. We evaluated different proteome extraction methods on root tips harvested from subsurface roots of naturally occurring *Populus* trees (Clinch River, East TN). Extraction of proteins via acetone precipitation worked best, and was employed towards proteomic profiling of a subset of *Populus* root samples obtained during PMI sampling trips to the Yadkin River in N. Carolina. Raw tandem mass spectra were searched against predicted proteins from genome sequences of *Populus trichocarpa* and *Laccaria bicolor*. Preliminary results show reproducible identification of >1000 proteins from each root sample, despite heterogeneity in morphology, differences in location, soil type, etc. Distributions of plant proteins across KOG (euKaryotic Orthologous Groups) functional categories were comparable across samples, suggesting similar protein metabolism in roots of sampled mature trees. Identified proteins occupied a variety of cellular locations (predicted by WolfPsort), proving the method to be efficient at extracting proteins in an unbiased manner. To date, >500 proteins were detected across all samples, forming an initial 'core' root proteome for natural *Populus* roots. Prominent in the core proteome were pectin methylesterases, major latex protein, peroxidases, enolase, glutathione S-transferase, alcohol dehydrogenase, ubiquitin, actin and histones. Relatively few *Laccaria* proteins were identified, with strong representation by histones and ubiquitin. Future studies on additional archived roots will provide information on whether observed proteome differences can be correlated with tree location, age, soil conditions, or other data acquired during sampling.

-Characterization of proteome changes resulting from RNAi knockdown of an auxin-related gene in *Populus We* have interrogated the root and shoot proteomes of wild-type (WT) *Populus*, and a strain with altered auxin signaling, IAA7.1 1-3 (see poster by Sukumar *et al.*). Initial measurements indicate that of >3900 proteins detected in shoots, 20 (28) were more (less) abundant in the IAA7.1 1-3 strain than in the WT. In roots, from >5500 proteins identified, 24 (1) proteins were more (less) abundant in the IAA7.1 1-3 strain. These measurements will potentially provide additional information on metabolic pathways affected by auxin signaling.

-Stable isotope labeling for comparisons via quantitative proteomics among multiple treatments, experimental conditions or time points Various quantitative proteomics methods have been developed, each with unique advantages. Label free quantification allows simultaneous protein identification and quantification without laboriously incorporating costly isotopes into samples. Metabolic labeling minimizes variability in sample preparation and measurement. Chemical labeling via iTRAQ/TMT permits multiplexed quantification. A comparison among these three techniques, using the latest-generation high performance mass spectrometer (LTQ-Orbitrap-Velos) provided guidance on selecting the most appropriate method for a proteomics study. The results indicate that iTRAQ/TMT chemical labeling has the highest quantification precision, label-free quantification provides the largest number of protein identifications, and metabolic labeling is intermediate in both measures. We have initiated experiments using iTRAQ/TMT quantification on experimental systems relevant to the PMI-SFA, including analysis of laboratory-grown *Populus* plants with and without inoculation of the roots with *L. bicolor*.

204

Plant-Microbe Interfaces: Isolation and Functional Characterization of Cultivatable Bacteria from the *Populus* Rhizo-Endosphere

Dale A. Pelletier* (pelletierda@ornl.gov), Tse-Yuan S Lu, Christopher W. Schadt, Neil Gottel, Matthew Foster, Marilyn Kerley, Patricia K. Lankford, and Mitchel J. Doktycz

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: Understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Populus trees are host to a variety of mutualistic microorganisms within their endosphere and rhizosphere that can have positive effects on the host. How these interactions manifest themselves within the spatially, structurally, and temporally complex scales of natural ecosystems is an open question. Our goal is to understand the diversity of the *Populus* microbiome and to elucidate the metabolic and molecular mechanisms responsible for shaping the *Populus*-microbial interface. In order to better understand the microbial communities associated with native *Populus deltoides* (Eastern cottonwood) we have under taken both cultivation independent and cultivation dependent assessment of *P. deltoides* rhizosphere and endosphere microbial communities. We have sampled *P. deltoides* at sites along the Caney Fork River in central Tennessee and Yadkin River in North Carolina. These sites represent ecotypes and soil conditions that are common to this region. We have sampled these sites in spring and fall to investigate seasonal changes in communities. This poster will focus on the microbial communities that were isolated from both rhizo- and endosphere sample utilizing direct plating methods. Isolated strains where identified by 16S rDNA sequence analysis and traits of interest in plant microbe interactions were investigated including plant colonization, both microbial and plant phenotyping and physiological properties.

A diverse array of bacterial strains (>1000) has been isolated from the rhizo- and endosphere of *Populus* roots. The isolates comprise 7 classes and 85 genera of bacteria including, Actinobacteria (14%), Bacilli (17%), Flavobacteria (6%), Sphingobacteria (3%), and a- (22%) b- (16%) and g- (22%) proteobacteria. Some general conclusions from isolation experiments are that native *Populus deltoides* roots are colonized by a diverse community of cultivable bacteria. The rhizosphere isolates are dominated by Actinobacteria and Bacilli strains, mainly *Streptomyces* and *Bacillus* species, the endophytes are dominated by a-proteobacteria while g-proteobacteria are prevalent in both environments. While we have cultivated a diverse array of organisms and our results are in general agreement with 454 16S pyrotyping results, the data suggests that specific groups of microbes may be especially underrepresented within the culture-based collection (e.g. *Flexibacter/Cytophaga/Bacteroides*, *Planctomycetales*, and *Acidobacteria*). A number of our isolates are

Pseudomonas species. This group of bacteria is known to have considerable genetic and phenotypic variability and is a common biocontrol and plant growth promoting bacteria, therefore more extensive characterization of 84 of these strains was performed. Full length 16S rDNA, rpoD and gyrB sequences indicate genotypic diversity of these isolates and phenotypic variability was found in a number of rhizosphere related traits including siderophore (95%), protease (51%), indole-3-acetic acid (50%) and phosphate solubilizing (41%) activities. Additionally root colonization of *Arabidopsis* and *Populus* identified plant growth promotion and modified root architecture phenotypes. To identify genes potentially responsible for plant phenotypes and rhizosphere colonization traits and comparative genome analysis we are sequencing the genomes of a number of isolates. A subset of these isolates are undergoing more extensive genomic and physiological analysis as well as plant-microbe co-culture experiments where plant physiology, transcriptome and metabolome are assayed which will lead to better understanding of the molecular mechanisms responsible for these interactions.

205

Plant-Microbe Interfaces: Bacterial and Fungal Communities Within the Roots and Rhizosphere of *Populus deltoides* in Upland and Lowland Soils

Neil Gottel,[1] Hector F. Castro Gonzalez,[1] M. Kerley, Zamin K. Yang,[1] Wellington Muchero,[1] Jessy L. Labbé,[1] Gregory Bonito,[2] Migun Shakya,[1,3] Dale A. Pelletier,[1,3] Mircea Podar,[1,3] Tatiana Karpinets,[1] Edward C. Uberbacher,[1,3] Gerald A. Tuskan,[1] Rytas Vilgalys,[2] Mitchel J. Doktycz,[1,3] and **Christopher W. Schadt**[1,3]* (schadtcw@ornl.gov)

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Biology Department, Duke University, Durham, N.C.; and [3]UT-ORNL Graduate School of Genome Science and Technology, Knoxville, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Populus trees are a genetically diverse and ecologically widespread riparian species, a potential cellulosic feedstock for biofuels, the first woody plant species to have a genome sequence, and are host to a wide variety of symbiotic microbial associations within their roots and rhizosphere. Thus they serve as an ideal model to study interactions between plants and microorganisms. However, most of our knowledge of microbial associations to date comes from greenhouse and young plantation-based trees; there have been no published efforts to comprehensively describe microbial communities of mature natural communities of *Populus*.

We compared root endophyte and rhizosphere samples collected at upland and lowland sites in Tennessee, to begin to understand the variation that might exist within and between soil types of these communities in both bacterial and fungal populations. 454 pyrosequencing was used to survey the microbial community of *P. deltoides*, using primers targeting the V4 region of the bacterial 16S rRNA gene and the D1 region of the fungal 28S rRNA gene. Further, genetic relatedness among the *Populus* trees was evaluated using 20 SSR markers chosen for distribution across all 19 linkage groups of the *Populus* genetic map. Jaccard's similarity coefficients were calculated based on segregation data generated from SSR marker-based assay. Soil physical, chemical and nutrient status, as well as tree growth and age characteristics were also evaluated.

121,540 bacterial and 322,100 fungal sequences were obtained, representing profiles of 20 endophyte and 20 corresponding rhizosphere samples. Bacterial rhizosphere communities were dominated by Acidobacteria (31%) and α-Proteobacteria (30%). Endophytic samples retained a lower proportion of α-Proteobacteria (23%), and were dominated by γ-Proteobacteria (54%). The fungal rhizosphere and endophyte samples all had equal amounts of the Pezizomycotina (40%), but differences were seen in the Agaricomycotina, which were more dominant in the rhizosphere (34%) than the roots (17%) and more specifically some endophytic root samples had, a large populations of a specific member of the Pucciniomycotina similar to known non-pathogenic basidiomycetous yeasts, however colonization of these OTUs was highly variable. Endophytic bacterial richness was also more variable and on average tenfold lower than the rhizosphere samples, suggesting root tissues provide a distinct environment supporting relatively few microbial types and that colonization events may be sporadic. Both fungal and bacterial rhizosphere samples showed distinct phylogenetic composition patterns compared to endophyte samples using UNIFRAC-PCoA analysis. PCoA analysis did not reveal changes in microbial communities between upland and lowland soil types in either rhizosphere or endophyte samples. Similarly, there was no strict adherence to soil type with regards to genetic similarity of sampled trees. In general, diversity among sampled *Populus* trees was lower among lowland genotypes with Jaccard's coefficients ranging from 0.02 - 0.08 compared to 0.08 - 0.13 for upland genotypes.

These findings indicate that the plant characteristics that influence the *Populus* root environment may represent a relatively stronger selective force than the soil environment in shaping the endophyte and rhizosphere microbial communities. However additional studies that carefully examine the variation in host genotype/phenotype vs. environmental effects will be required to fully describe these influences.

206

Plant-Microbe Interfaces: Acyl-homoserine Lactone Quorum Sensing in *Populus* Bacterial Communities

Amy L. Schaefer[1]* (amyschae@uw.edu), Joshua F. Emory,[3] **Dale A. Pelletier**,[2] Ryan Morlen,[1] Colin R. Lappala,[1] Bruce A. Tompkins,[3] Gary J. van Berkel,[3] E. Peter Greenberg,[1] and **Caroline S. Harwood**[1]

[1]Department of Microbiology, University of Washington, Seattle; [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [3]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: As part of our goal to understand the molecular and cellular events involved in establishing and maintaining beneficial interactions between *Populus* and its associated microbial communities, we are examining the role acyl-homoserine lactone signaling plays in this interaction. The long term project goals are to elucidate the AHL signal inventory of the *Populus*-associated bacterial community, define what these signals control in the bacteria that synthesize them as well as any effect AHLs may have on the plant host, as well as investigate how disruption of AHL signaling (either by genetic mutation or by addition of AHL-degrading enzymes) influences bacterial-plant interactions.

As part of the newly initiated ORNL Plant-Microbe Interfaces Science Focus area, we are characterizing the natural diversity of microbial associates of *Populus* and elucidating the molecular mechanisms by which these organisms interact. We sampled a population of *P. deltoides* as it occurs along the Caney Fork River in Tennessee in 2009. Analysis of 16s rDNA sequences indicates the *Populus* bacterial communities are dominated by Acidobacteria, Alphaproteobacteria, and Gammaproteobacteria (see poster by Gottel el al.) and the Proteobacteria are the predominant group isolated from *Populus* endophyte (86% of isolates, n=105) and rhizosphere (49% of isolates, n=157) samples (see poster by Gottel et al.). Many Proteobacteria use acyl-homoserine lactone (AHL) signals for cell density-dependent gene regulation, in a process known as quorum sensing and response. LuxI-type proteins synthesize small, diffusible AHL signals that function with LuxR-type signal receptors to control gene expression. Most known AHLs possess a fatty acyl side chain, derived from fatty acid biosynthesis, of varying side chain length and substitution. Recently we discovered that *Rhodopseudomonas palustris* makes a novel aryl signal, *p*-coumaroyl-HSL, which derives its side chain from an exogenously provided plant metabolite. This suggests that there may be additional novel HSL-type signals made by bacteria.

We screened 130 Proteobacteria isolated from *P. deltoides* for AHL production and found >80% Alphaproteobacteria and >20% of the Gammaproteobacteria isolates to be positive. This demonstrates that AHL signaling is prevalent in *Populus* microbial communities. We have also screened isolates for AHL production that is dependent upon exogenous addition of plant-derived aromatic compounds and found that *Enterobacter sp.* GM1 synthesizes cinnamoyl-HSL. This is the second aryl-HSL compound described to date and the first Gammaproteobacteria known to produce an aryl-HSL. We are sequencing the genome of GM1 to identify the cinnamoyl-HSL synthase and receptor genes. To enable a high-throughput means of detecting and identifying AHLs from culture supernatants, we developed a protocol that uses ultra high performance liquid chromatography (UPLC) coupled with multiple reaction monitoring (MRM) mass spectrometric detection. Using UPLC-MRM we can separate 21 synthetic AHL compounds, including the newly described aryl-HSLs, within four minutes and detect them at low concentrations (<0.2 pmol). AHL quorum sensing often controls the production of "public goods" such as antimicrobials and exoenzymes, as well as aggregation factors and conjugal transfer processes. In order to define the AHL regulon of a particular bacterium, mutants in either the *luxI*- or *luxR*-type genes are often constructed and analyzed relative to wild-type. However, not all AHL-producing bacteria are genetically tractable. To examine AHL-regulons in bacteria without constructing AHL-mutants we have demonstrated that purified AiiA lactonase, an enzyme that hydrolyzes the HSL ring of AHL signals, can be added to bacterial cultures to inhibit AHL-regulated phenotypes and gene expression. This protocol should enable future studies to define the AHL-regulons of *Populus*-associated bacteria.

207

Plant-Microbe Interfaces: The Effect of Host Species, Genotype, and Edaphic Factors on Rhizosphere Fungi Associated with *Populus deltoides*

Gregory Bonito[1]* (gmb2@duke.edu), Hannah Reynolds,[1] **Christopher W. Schadt**,[2] Jessy Labbé,[2] **Mitchel J. Doktycz**,[2] **Gerald A. Tuskan**,[2] and **Rytas Vilgalys**[1]

[1]Biology Department, Duke University, Durham, N.C.; and [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Populus deltoides is a common riparian tree species in areas of southeastern North America where regular flooding occurs. This species is reported to form both arbuscular and ecto- mycorrhizas in addition to harboring root endophytes. However, there is little understanding on the influence of edaphic or genotypic factors on the structuring of these rhizosphere fungal assemblages. To address the influences of genetic and soil factors on microbial root communities associated with *Populus* we designed a series of bioassay experiments using rooted cuttings to assess the diversity of fungi from different soils in an experimental greenhouse/growth chamber environment.

Specific objectives of this research were to:

1. Determine the influence of host species (*P. deltoides*, *Q. alba*, *P. taeda*) on the structuring of mycorrhizal communities
2. Determine the influence of *P. deltoides* genotype on the structuring of mycorrhizal communities
3. Compare the inoculum potential and influence of different soils on bacterial and fungal rhizosphere communities associated with a single *P. deltoides* genotype

Field soils from our ORNL *P. deltoides* research sites were used as the source of microbial inoculum in these bioassay studies. Cuttings from different *P. deltoides* genotypes were planted into a mixture of sterile sand and potting media amended with different field soils. Upon rooting, the cuttings form associations with fungi from different field soils, including mycorrhizal fungi. Experimental treatments included multiple field soils (inoculum) inoculated onto multiple *P. deltoides* clones. We also included a *P.deltoides x P.trichocarpa* hybrid, oak (*Quercus alba*) and pine (*Pinus taeda*) as positive controls and alternative hosts, as well as negative controls (no soil additions) in our experimental design. The plants were harvested after a five-month growing period. Soils were washed off the root systems, roots were visually assessed for ectomycorrhizas, and samples of bulk roots (representative of the whole root system) were taken from each plant for DNA extraction. Data on plant survivorship, number of shoots and shoot height was also recorded. Bulk root samples were then freeze-dried, pulverized, and DNA was extracted from them using a modified CTAB-chloroform extraction protocol. A number of different primer sets were tested and a subset of these was selected to amplify targeted microbial groups for pyrosequencing. The fungal community from each root will be sampled for both Fungi were ITS and LSU rDNA regions using the fungal specific primers ITS1f & ITS4 and LROR & LR3. Arbuscular mycorrhizae were amplified selectively using the primer set AML1 & AML2. Bacterial 16S rDNA primers that discriminate against plastid DNA were used to compare rhizosphere bacterial communities in selected samples. We made clone libraries from amplicon pools using these primer sets. From these sequences we have assessed these primer sets are efficient at amplifying the targeted groups. From our visual assessments, pine and oak seedlings had high ectomycorrhizal colonization (>80%) while most of the *P. deltoides* genotypes appeared to have generally low ectomycorrhizal colonization (<10%). One exception was the D124 genotype, which had the highest rates of ectomycorrhizal colonization of any of the *Populus* genotypes tested. Ectomycorrhizal species in the genus *Tuber*, *Hebeloma*, and *Laccaria*, were recovered from both oak and *Populus* roots. Further, one of the species of *Tuber* recovered from *Populus* roots in our bioassay was identical to the sequence for the only ectomycorrhizal fruitbodies (of truffles) that have found fruiting in our *Populus* field plots so far. We have collected this undescribed truffle species (*Tuber*) under *P. deltoides* at our ORNL sites in Tennessee and North Carolina. From our arbuscular mycorrhizae clone libraries we recovered two species of *Paraglomus*, one species of *Gomus*, and a novel species that falls in-between the currently known families within the Glomeromycota. We are now in the process of single direction pyrosequencing using 454 Titanium Lib L chemistry. Results from these experiments will be presented.

208

Plant-Microbe Interfaces: Identification of Quantitative Trait Loci and Targeting of Genes Affecting Ectomycorrhizal Symbiosis in *Populus*

Jessy Labbé[1]* (labbejj@ornl.gov), Wellington Muchero,[1] Lee E. Gunter,[1] Annegret Kohler,[2] Véronique Jorge,[3] Catherine Bastien,[3] Francis Martin,[2] François Le Tacon,[2] **Gerald A. Tuskan**,[1] and **Mitchel J. Doktycz**[1]

[1]BioSciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]INRA, Nancy, Champenoux, France; and [3]INRA, Orléans, France

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

The tree rhizosphere hosts a large community of microbes that compete and interact with each other and with plant roots. Within this community of microorganisms, ectomycorrhizal fungi are almost ubiquitous. Mycelium of symbiotic fungi and root tips form a novel composite organ, so-called ectomycorrhiza, which is the site of nutrient transfer between the symbionts. Because genome sequence is available for both *Populus trichocarpa* and the basidiomycete *Laccaria bicolor*, the *Populus-Laccaria* symbionts are an excellent system to study ectomycorrhizal interactions.

We have analyzed a *Populus deltoides* × *P. trichocarpa* F_1 pedigree (Family 54B, INRA-Orléans, France) for quantitative trait loci (QTLs) affecting ectomycorrhizal development and for microarray characterization of gene networks involved in this symbiosis. A 300 genotype progeny set was evaluated for its ability to form ectomycorrhiza with *L. bicolor*. The percentage of mycorrhizal root tips was determined on the root systems of all 300 progeny and their two parents. QTL analysis identified four significant QTLs, one on the *P. deltoides* and three on the *P. trichocarpa* unsaturated genetic maps (Jorge et al. 2005). These QTLs were aligned to the *P. trichocarpa* genome and each contained several megabases and encompass numerous genes. Using cDNA from RNA extracts of ectomycorrhizal root tips from the parental genotypes *P. trichocarpa* and *P. deltoides*, expression analysis from a NimbleGen whole-genome microarray, was used to narrow the candidate gene list. About 3.4% of the *Populus* gene models were differentially expressed (1,543 genes; p-value≤0.05; ≥5.0-fold change in transcript level) in mycorrhiza of the two parents including genes coding for the lignin metabolism and the NBS-LRR class of disease resistance proteins. Forty-one transcripts were located in the QTL intervals. Among these 41 transcripts, 25 were overrepresented in *P. deltoides* relative to *P. trichocarpa*; 16 were overrepresented in *P. trichocarpa*. The transcript showing the highest overrepresentation in *P. trichocarpa* mycorrhiza libraries compared to *P. deltoides* mycorrhiza codes for an ethylene-sensitive EREBP-4 protein that may repress defense mechanisms in *P. trichocarpa* while the highest overrepresented transcripts in *P. deltoides* code for proteins/genes typically associated with pathogen resistance. Recently we genotyped 300 mapping progeny on a 6K Illumina *Populus* SNP array to improve the genetic maps and increase accuracy of gene targeting. Finally, these results suggest that there is a shared molecular communication network between these two organisms and that modification of metabolic pathways may be occurring before, during and after colonization.

209

Plant-Microbe Interfaces: The Discovery of Novel Secretory Motifs Modulating Plant-Microbe Interactions

Ting Li[1]* (lit1@ornl.gov), Jennifer L. Morrell-Falvey,[1] Gregory B. Hurst,[2] Timothy J. Tschaplinski,[1] **Gerald A. Tuskan**,[1] and **Xiaohan Yang**[1]

[1]Biosciences Division and [2]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Recent data has shown that proteins less than 200 amino acids in length that are encoded in short open reading frames (i.e., small proteins) can modulate diverse biological processes, including signal transduction between plants and their bacterial or fungal associates. The goal of this effort is to systematically investigate the functional genomics of the small signaling proteins mediating plant-microbe interactions.

It is widely accepted that small proteins play important roles in plant growth and development, such as transcriptional regulation, signal transduction, stress response and defense response. Transcriptomic analyses in *Populus* revealed thousands of short open reading frames expressed under normal and drought conditions and putative small signaling proteins were identified by additional comparative genomics analysis. Despite these efforts, the prediction and annotation of small proteins remain challenging. We report here a computational approach to predict small signaling proteins mediating plant-microbe interactions using protein signatures.

We hypothesize that novel conserved protein domains/motifs are signatures representing the functions of small proteins. A large-scale analysis of the conserved domains in small proteins from five plant species, including *Populus trichocarpa*, *Vitis vinifera*, *Arabidopsis thaliana*, *Cucumis sativus* and *Glycine max*, was performed. We first identified conserved protein domains using sequence-based probabilistic models, then known protein motifs were removed by querying 14 current protein domain databases. Our analysis identified 732 motifs that are not documented in the public protein domain databases. Most of these novel motifs are over-represented in small proteins and a larger portion of these motifs are located in the N- or C-terminus of the small protein sequences compared to known motifs. In addition, we found a distinctive expression pattern for the small proteins containing novel motifs as compared with those containing known domains. A significantly higher percentage of the small proteins containing the novel motifs, relative to those of the known domains, were predicted to locate in the extracellular space, suggesting that some of these novel protein motifs may be signatures for protein secretion or intercellular signaling. Computational and experimental characterizations are underway to determine the potential functions of these novel protein motifs in plant-microbe interactions. The novel motifs uncovered in this research will facilitate the genome-wide discovery of small proteins functioning in intercellular signaling in plant species.

210

Plant-Microbe Interfaces: Transcript and Metabolic Networks Underlying Induced Systemic Resistance in *Arabidopsis* Co-cultured with *Pseudomonas* Strains

Abhijit A. Karve[1]* (karveaa@ornl.gov), Sara M. Allen,[1] Sara S. Jawdy,[1] Lee E. Gunter,[1] Nancy L. Engle,[1] Timothy J. Tschaplinski,[1] Tse-Yuan S. Lu,[1] Dale A. Pelletier,[1] Jennifer L. Morrell-Falvey,[1] Jay Chen,[1] Andrey A. Gorin,[2] Nikita D. Arnold,[2] Tamah Fridman,[2] Gerald A. Tuskan,[1] Mitchel J. Doktycz,[1] and **David J. Weston**[1]

[1]Biosciences Division and [2]Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

In addition to providing support, nutrients and water, plant roots also act as communication conduits with soil microflora. Such plant-microbe interactions can elicit an array of beneficial or unfavorable phenotypes via systemic signaling. In the case of rhizobacteria including some *Pseudomonas* strains, plant growth is promoted by the suppression of diseases and insect herbivory. This phenotype termed as induced systemic resistance (ISR) is mediated through complex metabolic and hormonal networks. This is phenotypically similar to the pathogen-induced systemic acquired resistance (SAR) except that SAR is dependent on the accumulation of salicylic acid (SA) and pathogenesis related (PR) proteins. Alternatively, ISR relies on jasmonic acid (JA) and ethylene signaling in *Arabidopsis*. Our recent discovery of a putative *Pseudomonas* strain (GM-30) from the rhizosphere of *Populus deltoides* brings to question whether this strain elicits a novel beneficial, neutral or antagonistic plant phenotype. Here, we report the initial characterization of the plant systemic response induced from co-cultures with *Pseudomonas* strain*s* GM-30 and Pf-5. Root colonization of *Arabidopsis* by both Pf-5 and GM30 promoted plant growth and modified plant root architecture. Transcript profiles were collected from root and shoot tissue on days 3 and 10 after inoculation. Ontology analysis of microarray data suggests that jasmonic acid and ethylene biosynthesis were induced in shoots of plants co-cultured with Pf-5 relative to those co-cultured with GM30. Furthermore, plants co-cultured with GM30 showed significant enrichment (Wilcoxon rank sum, p = 0.007) for genes encoding PR proteins. These results suggest that both *Pseudomonas* strains trigger plant systemic defense but through alternate pathways. To confirm that both *Pseudomonas* strains colonize the plant roots, fluorescence in situ hybridization (FISH) was performed using a probe specific for γ-Proteobacteria (GAM42a-Alexa488) and the universal bacterial probe EUB338 (EUB388-Alexa594). These data indicate that the bacteria were physically associated with and formed colonies on the roots. To validate our array results, a qPCR pathway index populated with 42 published marker genes associate with both beneficial and antagonist plant-microbe interactions was created using a high throughput qPCR platform. Results from this analysis further suggest that Pf-5 elicits ISR through JA and ethylene signaling, while GM-30 altered the expression of genes associated with SA biosynthesis and PR genes. Metabolite profiles of plant shoots at day 3 after inoculation found that Pf-5 co-cultures induced methionine, a precursor for ethylene signaling relative to GM-30 co-cultures. Shoots from plants co-cultured with GM-30 had higher levels of phenylalanine and tryptophan, which are precursors for SA and auxin biosynthesis, respectively. Taken together, these results suggest that Pf-5 and GM-30 play a role in ISR and SAR. This hypothesis is currently being tested with a *Pseudomonas syringe* challenge on *Arabidopsis* seedlings co-cultured with either Pf-5 or GM-30. Current research is being conducted to compare the *Arabidopsis* and *Populus* systemic responses when co-cultured with Pf-5 and GM-30 by using comparative network analyses for genes, metabolites and proteins.

Protein network comparisons require comprehensive characterization of the proteomes in the co-cultivated and control samples, as incomplete or false protein identifications are recognized as the source of incorrect conclusions about induced biological changes. We are developing novel mathematical and computational algorithms for MS/MS-based proteomics as a promising way to increase precision and improve reproducibility of proteome profiles. So far new approaches were developed and tested on two key stages of the pipeline that transforms raw MS/MS spectra into protein ids: (1) assignment of MS/MS spectra to peptides; and (2) conversion of reliably assigned peptides to proteins.

At the peptide identification stage rigorous analytical formulas were used to estimate confidence of each assignment rather than usual empirical or semi-empirical generic cutoff approaches. The established confidence values were integrated for calculations of protein reliability by Bayesian statistics. In the preliminary trials we observed an increase of 2-3 times in the number of useable peptides and 1.25-1.5 boosts in the number of unique protein ids. The ongoing experiments will evaluate effects of these advances on the reproducibility and dynamic range of proteome measurements.

211

Plant-Microbe Interfaces: Differential Involvement of Auxin Signaling Components in Modulating Root Responses to Various Rhizosphere Microbes

Poornima Sukumar[1]* (sukumarp@ornl.gov), Alyssa Delong,[2] Whitney McNutt,[1] Jennifer Morrell-Falvey,[1] David J. Weston,[1] Dale A. Pelletier,[1] Karuna Chourey,[1] Gregory B. Hurst,[1] Valérie Legué,[3] Annegret Kohler,[3] Francis Martin,[3] Mitchel J. Doktycz,[1] Gerald A. Tuskan,[1] and **Udaya C. Kalluri**[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Purdue University, West Lafayette, Ind.; and [3]INRA, Nancy, Champenoux, France

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Understanding the mechanism of plant-microbe interactions is pertinent as microbes can influence plant growth, positively or negatively. Many mechanisms of interaction have been postulated including the involvement of phytohormones such as auxin. Several microbes in the rhizosphere have been shown to produce auxin and have been shown to affect root architecture including changes in primary and lateral root length, number of laterals, production of tertiary roots and nodulation. This study looks at the influence of three different microbes; *Piriformospora indica*, *Laccaria bicolor* strain S238N and a *Pseudomonas* strain GM30 on root architecture of *Arabidopsis* and *Populus*. 5-day-old *Arabidopsis* seedlings were co-cultured with the listed microbes and the root architectural modifications were examined 7 days later. We find that the 3 organisms used in this study have different affects on plant roots. While *P. indica* enhanced the number of secondary roots that emerge, *L. bicolor* co-culture produced longer laterals and GM30 enhanced the density of lateral root formation. In *Populus* tissue culture plants, there was an increase in the development of secondary adventitious roots with *L bicolor* and *P. indica*. Exogenous auxin application mimicked the phenotype observed with some of the strains. Additionally, there was enhanced expression of AtGH3-GUS, an auxin induced reporter line with GM30 co-culture, in the roots, indicating that this microbe could alter auxin accumulation in roots. The inhibition of lateral root formation through local application of auxin transport inhibitor can be compensated by *P. indica* and GM30, and to a smaller extent by *L. bicolor* co-culture. This supports previous reports that auxin is an important modulator of microbe-altered root architecture. To genetically dissect the components of auxin signaling, and transport involved in modulating various architectural phenotypes, we examined the root phenotypes of mutants available in *Arabidopsis* and a few AUX/IAA RNAi mutants in *Populus*, with co-culture. Among the signaling mutants, we find that *iaa34, iaa7* and *gh3-17* display altered response *to P. indica,* while *arf7* and *arf19* display altered response to GM30. Interestingly, *iaa3* displays altered response to all the 3 microbes tested in this study. The *Populus* mutant *iaa7* also seemed to have reduced responses to *L. bicolor* and *P. indica*. Among the mutants defective in auxin transport, we find that *aux1-7, lax2, lax3, pin3, pin7, pin3-pin7* and *abcb4-1* display altered response to *P. indica* and *L. bicolor*. These results suggest that the different microbes may use alternate components of the signaling and transport pathway of auxin to produce respective architectural alterations. To see if the identified auxin signaling, and transport proteins change with co-culture, RT-PCR experiments are underway to understand changes of respective signaling and transport genes over the duration of the experiment. Additionally, transgenic plants that have GFP/YFP fused with auxin transport proteins are being used to determine if the localization and amount of these proteins change over time. We find that PIN2-GFP and AUX1-YFP expression reduce with *P. indica* and GM30 at 5-6 days after infection. The accumulation of flavonoids, a class of secondary metabolites that have been shown to inhibit auxin movement, was also found to be altered in the presence of the tested microbes. Together, these results indicate that the alteration of root architecture during plant-microbe interaction involves complex regulation of auxin transport and signaling, possibly unique to individual microbes.

Student Oral Presentation–Tuesday

212

Plant-Microbe Interfaces: Experimental and Computational Approaches for Microbial Diversity Characterization Using Artificial Communities

Migun Shakya[1,2*] (shakyam@ornl.gov), Zamin K. Yang,[2] **Christopher W. Schadt**,[2] and **Mircea Podar**[2]

[1]UT-ORNL Graduate School of Genome Science and Technology, Knoxville Tenn.; and [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

The use of 16S ribosomal RNA gene sequencing (rDNA) to characterize the taxonomic diversity and abundance of organisms plays an important role in microbial ecology studies. Microbial diversity characterization by pyrosequencing dramatically increases the scale at which environmental samples can be analyzed both in number and sequence depth. Nevertheless, the approach often has limitations for which it is difficult to control or account. Such limitations include unequal efficiency of DNA isolation, unknown ribosomal gene copy number, biases in DNA amplification, and sequencing errors. In addition, the myriad of computational tools available to analyze the data makes the data analysis even more confusing. Previous studies aimed at improving the techniques and overcoming some of those limitations were limited by sequencing depth, as well as the diversity of completed reference genomes. To determine the effects of experimental and computational steps involved in characterization of microbial diversity by 454 sequencing, we constructed artificial genomic DNA communities using cultivated organisms that have known genomic sequence and rDNA copy number. More than 60 species representing 14 bacterial and 3 archival phyla were included in the analysis. Genomic DNAs were mixed at known concentrations and hypervariable regions of the 16S rDNA gene (V1-2, V1-3, V4, V3-5, V3-9, and V6-9 for bacteria and V4, V3-9 for archaea) were amplified and sequenced with the 454 FLX and 454 Titanium system. In order to test the primer bias we also carried out a metagenomic study of the artificial community. The data was analyzed in terms of PCR/sequencing errors, chimeras, number of OTUs, abundance of individual alleles and species. For many taxa, the inferred abundance matched relatively well to the composition of the assembled communities. However, significant primer-dependent biases were observed for particular species or even phyla. A single set of 16S rDNA primers may, therefore, incompletely represent the diversity present in natural microbial communities. Similarly, the choice of data analysis pipelines also produced different results terms in number of OTUs. The computational tools should, thus, be carefully chosen based on the posed question. In addition, mock communities with known members should be constantly tested to authenticate the sequencing protocol and the computational tools.

213

Plant-Microbe Interfaces: Emerging Technologies for the Functional Characterization of Isolates from *Populus*

Jennifer Morrell-Falvey[1]* (morelljl1@ornl.gov), Suresh R. Neethirajan, A. Nicole Edwards,[1,2] Bernadeta Srijanto,[3] Roy Dar,[3] Sarah Melton,[1] Dale Pelletier,[1] **Scott T. Retterer**,[1,3] and **Mitchel J. Doktycz**[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]UT-ORNL Graduate School of Genome Science and Technology, Knoxville, Tenn.; and [3]Center for Nanophase Materials Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

The composition of microbial communities within and around plant species is dependent on dynamic physical and chemical signaling events that occur within the local environment and at the root surface. Visualization and quantification of these events in natural systems is challenging. However, emerging technologies that combine advances in nanostructure fabrication, microfluidics and imaging provide a means of recreating these events within model systems. These systems mimic aspects of their natural counterparts while providing tractable experimental platforms in which both individual cellular responses and population dynamics can be recorded and analyzed.

Model systems, amenable to imaging, that allow dynamic modulation of local physicochemical cues in a controllable manner have been developed to recreate the interactions between microbes and their hosts. Building from work aimed at sampling and cultivating isolates from the *Populus* rhizo- and endospheres these tools will provide a means of screening the chemotactic response, surface adherence, and colonization dynamics of individual *Populus* isolates.

A nanostructured microfluidic platform has been created in order to examine the chemotactic responses of isolates to specific plant-associated signals. This platform is created using a combination of electron beam lithography and anisotropic silicon etching techniques. It can be easily replicated via silicone molding and facilitates the physical tracking of hundreds of microbes within a quasi-two-dimensional space that confines microbes within the focal volume of a conventional phase contrast microscope without significantly impeding natural motility. A nanostructured interface or membrane separates the main "chemotaxis" channel from two "feeder" channels that allow spatiotemporal modulation of the chemical environment within the main channel with negligible hydrodynamic disturbance of the microbial population (Figure 1).

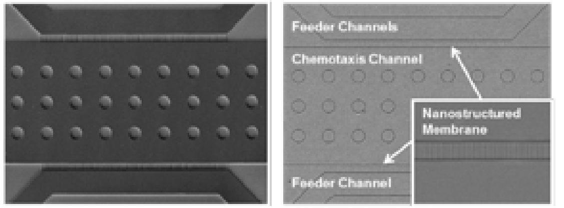


Figure 1. (A) A silicon and polymer master is formed using a combination of electron beam lithography, anisotropic silicon etching and crosslinkable polymer. (B) This master can be replicated in silicone to form a microfluidic chamber with nanostructured membranes that allow material exchange during real-time imaging of the microbial response

The dynamics of material exchange within the system has been characterized using fluorescence microscopy. Complete modulation of the local chemical environment within the device can occur within a ten-minute cycle, allowing changes in microbial motility to be monitored over time. Additionally, administration of multiple reagents from separate feeder channels enables the creation of chemical gradients in which population dynamics can be monitored. Proof-of-concept studies have been carried out, examining differences in motility and chemotactic response in *A. brasilense* and three Che1 mutant strains. Differences in velocity, reversal frequency and rate of directional change were recorded and quantified under different environmental conditions. Additional studies are being conducted to examine the response of *Populus* isolates to known plant metabolites.

Imaging studies of colonization and surface adherence have been carried out using confocal fluorescence imaging and atomic force microscopy. Real-time, 3-dimensional imaging of colonization dynamics was carried out in *Populus* roots using natural isolates, transformed to express GFP. The combined autofluorescence from the plant roots and GFP expression from the isolates allowed the growth of microbial colonies within and around the roots to be tracked over time. Atomic force microscopy was used to track the evolution of microbial biofilms from *Populus* isolate with even greater resolution. AFM analysis enabled the observation of pili formation and the evolution of distinct microbial morphologies over the course of the biofilm formation.

Taken together, the use of emerging technologies for imaging and the creation of model systems allow the observation and quantification of microbial responses to specific changes in their environment at scales that are unprecedented in natural systems. Moving forward with these technologies, we look towards the observation of complex microbial communities to better understand community dynamics across the plant-microbe interface.

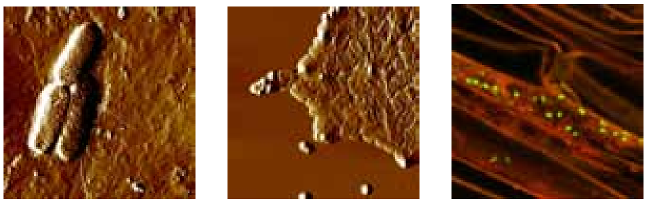


Figure 2. (A & B) Protocols for preserving pili and ultrastructural components of bacterial isolates for atomic force microscopy have been developed. Imaging on colonies of GM30 and YR343 strains on biofilms grown for 4,8,16 and 24 hours on mica was executed to better understand the expression of pili and fimbriae, as well as overall biofilm morphology. (C) *Populus* roots were fixed in 4% paraformaldehyde for detection of associated bacteria by fluorescence in situ hybridization (FISH). Bacterial probe EUB338 was labeled with Alexa488 (green). Plant roots are detected by autofluorescence (red).

214

Plant-Microbe Interfaces: Collaboration Platform for Scientific Communication, Management, Information Storage and Sharing

Guruprasad Kora[2]* (koragh@ornl.gov), Michael Leuze,[2] Tatiana Karpinets,[1] Mustafa Syed,[1] Susan Holladay,[1] **Edward C. Uberbacher**,[1] and **Mitchel J. Doktycz**[1]

[1]Biosciences Division and [2]Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Plant-Microbe Interfaces (PMI) knowledgebase represents a unique platform for biologists, analysts and data collectors to share, collaborate and analyze data from a single point of access. Its goal is to consolidate all team related low-level scientific data as well as document-level project related information into an interactive computational environment. The PMI knowledgebase environment offers exceptional computational and collaborative capabilities to biologists with little or no computational background. The PMI portal is currently serving 80 members of the Department of Energy (DOE) funded Plant Microbe Interfaces scientific focus area. The PMI knowledgebase comprises of three integrated components: a data sharing and collaboration platform, a Laboratory Information Management System (LIMS) interface, and a content management system with built-in querying engine. Each of these components is presented with an easy to use web interface with appropriate security infrastructure built-in.

The PMI portal's collaborative platform was designed from the ground up, with scientific team collaboration and data sharing aspects in mind. It utilizes modern collaboration and social networking features that provide structure-less social utility tools that connect and facilitate a group of like-minded co-workers to share, collaborate and discuss on a given scientific task. Furthermore, the platform enables social features such as "Scientific Walls" to provide single point discussion threads to facilitate better inter-group interaction. This approach towards team collaboration turns out to be the best technique to assimilate and disperse data-driven knowledgebases.

The PMI portal's LIMS interface enables users to access the underlying data management layer with an easy to use and intuitive web interface. The interface seamlessly connects to laboratory-wide LIMS environment and makes day-to-day tasks like raw-data fetching and data summarization extremely efficient and easy. Furthermore, the portal provides a dynamic data analytics environment that facilitates users to perform standard statistical analysis on the LIMS stored data.

The portals content management layer handles document-level data across portal users. It provides a single unified repository to manage any content type—documents, images, data sheets, archives etc. The content management system has an intelligent reference engine that enables real-time content tagging with project entities like goals, project-wide events, and system participants, thus enabling intelligent tracking of document life-cycle. The content management system is fine-tuned towards scientific data management and retrieval processes as it provides document preview features to scientific data types directly within your browser, without having to download them. The system provides a powerful search capability that enables in-document and free-text searching across all managed content. An alert system works in conjunction with the document management system to enable real-time tracking of content across the system.

PMI knowledgebase and Portal can be accessed at *http://pmi.ornl.gov*

215

Plant-Microbe Interfaces: Exploring Mutualistic and Parasitic Symbiotic Relationships of Microbes With Plants Using PMI Knowledgebase Tools

Tatiana Karpinets[1]* (karpinetstv@ornl.gov), Byung Park,[2] Michael Leuze,[2] Guruprasad Kora,[2] Mustafa Syed,[1] Dale Pelletier,[1] Christopher Schadt,[1] **Edward Uberbacher**,[1] and **Mitchel J. Doktycz**[1]

[1]Biosciences Division and [2]Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://PMI.ornl.gov

Project Goals: The goal of the Plant-Microbe Interfaces science focus area is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as an initial test system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we will focus on 1) characterizing the natural variation in *Populus* microbial communities within complex environments, 2) elucidating *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships, and 3) performing metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Symbiotic interactions between microbial organisms and plants can be mutualistic, when it is beneficial for both of them; parasitic, when the microbial organism benefits at the expense of the plant; or commensal, when the microbial organism benefits without damaging the plant. Although some cases of plant-microbe interactions are ambiguous and cannot be strictly classified as beneficial, parasitic, or commensal, many microbial organisms are well-known pathogens of plants and cause devastating diseases to their host. The other interactions, like mycorrhizal associations between plant roots and fungi or associations of nitrogen-fixing rhizobium bacteria with legumes, are well-known examples of resource-resource based mutualism, when one type of resource produced by the microorganism is traded for another resource produced by the plant. The distinct mutualistic and parasitic phenotypes of microorganisms make their genomes a valuable target for comparative analyses. Each of these symbiotic relationships involves adaptation and evolution of microbes and leads to appropriate changes, not only in the phenotypes of the organisms, but also in their genomes. Thus, specific molecular functions, metabolic pathways and biological processes should exist in microbial genomes that underlie the mutualistic or parasitic nature of symbionts and, likely, distinguish one phenotype from the other.

In this study we used a comparative analysis of bacterial genomes representing different types of plant-bacterial symbionts to explore genomic features underlying mutualistic and parasitic bacterial phenotypes. These phenotypes will be referred as plant pathogens and plant endophytes. An initial challenge of the study was to compile a comprehensive set of sequenced pathogens and endophytes for the comparison. This was addressed by developing a novel tool, the "Genes/genomes On-Line explorer" (GOLexplorer) as a part of the PMI knowledgebase. The tool allows one to infer confident relationships between characteristics of genes or organisms if their classifications in terms of these characteristics are available for a set of the objects. If the object is an organism, for example, the characteristics can include a type of symbiotic relationship, a type of metabolism, its host, a taxonomic group, or preferred temperature range. The developed tool uses the association rule-learning algorithm to find confident relationships between such characteristics. We have used GOLexplorer to discover associations between characteristics of bacterial organisms available in the GOLD database and to compile representative lists of plant pathogens and endophytes for further genome comparative analyses.

Our initial selection of plant symbionts using the GOLexplorer identified 28 sequenced endophytes and 36 pathogens with no significant differences between the groups in genome sizes, GC contents and taxonomic classification of the organisms at the level of phylum and order. In both groups most organisms (~80%) belonged to the phylum of proteobacteria. At the level of order, both, pathogens and endophytes, have representatives of Acholeplasmatales, Burkholderiales, Enterobacteriales, Pseudomonadales, Rhizobiales, and Xanthomonadales. Most sequenced endophytes (90%) were also annotated by the "nitrogen fixation" phenotype, but only one pathogen had this annotation. Species in the dominating orders of pathogens often had several representatives of the same genus and even of the same species. To equal the number of organisms in both groups we have removed such duplicates. This filtering resulted in 28 pathogens and 28 endophytes for further comparative analysis at the level enzymes and metabolic pathways using the Pathway Tools available in the PMIcyc and our recently developed toolkit for prediction in the genomes of Carbohydrate-Active enzymes (CAZymes).

One interesting finding of the analysis was the presence of distinct enzymatic signatures for pathogens and endophytes in terms of enriched CAZy families in their genomes. The genomes of all endophytes were enriched in CAZymes with a constructive (biosynthetic) metabolic activity. The enzymes belonged to several families of glycosyltransferases involved in the synthesis of oligo- and polysaccharides. Genomes of pathogens had a relative abundance of CAZymes with a destructive (degrading) metabolic activity. These enzymes belonged to the family of glycosyl hydrolases, which are involved in the release of glucose from oligo- and polysaccharides. This set of enzymes was more specific for each pathogenic organism, most likely, because different sugars can dominate in different plant hosts. These and other findings will be presented, which describe the different metabolic profiles of endophytes and pathogens.

The Predictive Microbial Biology Consortium

216

Characterization of Naturally Occurring and Model Microbial Communities

Jennifer J. Mosher,[1] James G. Moberly,[1] Christopher W. Schadt,[1] Tommy J. Phelps,[1] Mircea Podar,[1] Steven D. Brown,[1] Anthony V. Palumbo,[1] Michael W.W. Adams,[2] David A. Stahl,[3] Kristina L. Hillesland,[3] Judy D. Wall,[4] Matthew W. Fields,[5] Terry C. Hazen,[6] and Dwayne A. Elias[1]* (eliasda@ornl.gov)

[1]Oak Ridge National Laboratory; [2]University of Georgia; [3]University of Seattle; [4]University of Missouri; [5]Montana State University; and [6]Lawrence Berkeley National Laboratory

Project Goals: Microbial community construction is ongoing with *Desulfovibrio vulgaris*, *Geobacter sulfurreducens* and *Methanococcus maripaludis* to determine carbon mineralization, energy balance and electron accepting patterns in model communities. Cultures are assessed as syntrophic and competitive communities to examine intercellular communication at several omic levels and is in coordination with *D. vulgaris*/*M. maripaludis* co-culture and evolution studies, stress conditions implemented in-situ at Hanford, and utilizing *D. vulgaris* mutants to determine particular gene importance on community function. All data will help construct metabolic models at different complexities with other ENIGMA groups. The aims of in-situ microbial community assessment in Hanford groundwater are to determine the temporal population succession while isolating new keystone species. Further, these approaches can be used as a proxy for in-situ tests in an effort to predict the results of different perturbations.

Microbial community construction, cultivation and analyses are being performed using the metal-reducing bacteria *Desulfovibrio vulgaris* Hildenborough and *Geobacter sulfurreducens* PCA as well as *Methanococcus maripaludis* S2 to determine complete carbon mineralization, energy balance and electron accepting patterns in consortia communities. Mono-, co-, and tri-cultures are being assessed both as syntrophic and competitive communities to examine cell to cell communication at several omic levels and are being performed in coordination with *D. vulgaris*/*M. maripaludis* coculture and evolution studies (D. Stahl and K. Hillesland). These communities will be subjected to stress conditions being implemented *in-situ* at Hanford (T. Hazen) while also utilizing several *D. vulgaris* mutants to determine the importance of particular genes and metalloproteins on community function (with J. Wall and M. Adams). Technologies developed include multispecies microarrays, along with species specific qPCR primers and fluorescent antibodies to better understand intercellular coordination within these defined communities. All data will help construct metabolic models at different complexities with microarray data being directly comparable to existing datasets from other ENIGMA groups. Future ENIGMA efforts are likely to be able to take advantage of similar technologies and these can be developed quickly for new keystone organisms of interest.

The aims of *in-situ* microbial communities assessment in Hanford groundwater are to determine the temporal population succession while isolating new species key to the function of these populations. Recently, Hanford 100H groundwater (from T. Hazen) was inoculated into triplicate, custom designed flow through reactors and incubated (30°C) for 95 days with a 100 hour generation rate. Protein, gas and liquid metabolite quantification, and 16S rDNA identification of the microbial community members were highly reproducible with the final community dominated by the genera *Pelosinus*, *Acetobacterium*, *Methanobacterium* and *Methanosarcina*. Six new genera and seven new species of sulfate- and Fe(III)- reducing bacteria were isolated including three new *Pelosinus* spp. All isolates are currently being assessed for Fe(III), U(VI), and Cr(VII) reduction. In collaboration with Mike Adams (U. Georgia) preliminary results show that cell uptake rates of V, Fe, Co, Se, W, and Mo are coordinated to the temporal succession and recession of different community groups. The follow-on experiment is to repeat these conditions with and without *in-situ* Cr(VII) levels to determine the Cr influence on planktonic community structure, in coordination with biofilm studies (M. Fields, Montana State U.). Short and longer-term environmental perturbations will be coordinated with recently developed ENIGMA field plan for Hanford to test hypotheses developed from the *in-situ* experiments. Such efforts and integration will generate a more comprehensive understanding of the community and reaction to perturbations, while supplying new microbial consortia and isolates to the wider ENIGMA team in order to further ENIGMA and DOE goals.

217

Student Oral Presentation–Monday

Characterizing the Metalloproteomes of Model Microorganisms

W. Andrew Lancaster[1]* (lancaste@uga.edu), Angeli L. Menon,[1] Sunil Kumar,[1] Farris L. Poole,[1] Ming Dong,[2] Megan Choi,[2] Mark Biggin,[2] Haichuan Liu,[3] H. Ewa Witkowska,[3] Sunia A. Trauger,[4] Gary Siuzdak,[4] Steven M. Yannone,[2] John A. Tainer,[2,4] and **Michael W.W. Adams**[1]

[1]University of Georgia, Athens; [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [3]University of California, San Francisco; and [4]Scripps Research Institute, La Jolla, Calif.

Project Goals: Metal ion co-factors afford proteins virtually unlimited catalytic potential, enable electron transfer reactions and greatly impact protein stability. Consequently, metalloproteins (MPs) play key roles in virtually all biological processes. However, predicting the types of metal that an organism utilizes in its metalloproteome from its genome sequence is currently impossible since metal coordination sites are diverse and poorly recognized. Determining the identity of MPs directly from native biomass can resolve some of these issues. We are using *Pyrococcus furiosus*, a hyperthermophilic archaeon that grows optimally at 100°C, as the model organism.

Metal ion co-factors afford proteins virtually unlimited catalytic potential, enable electron transfer reactions and greatly impact protein stability. Consequently, metalloproteins (MPs) play key roles in virtually all biological processes. However, predicting the types of metal that an organism utilizes in its metalloproteome from its genome sequence is currently impossible since metal coordination sites are diverse and poorly recognized. Determining the identity of MPs directly from native biomass can resolve some of these issues. We are using *Pyrococcus furiosus*, a hyperthermophilic archaeon that grows optimally at 100°C, as the model organism. Large scale fractionation of native biomass using non-denaturing, sequential liquid chromatography (26 columns) coupled with high-throughput tandem mass spectrometry (HT-MS) to separate and identify proteins led to the identification of ~80% (967) of the cytoplasmic proteins in *P. furiosus*. By coupling native biomass fractionation with inductively coupled plasma mass spectrometry (ICP-MS), a robust, metal-based approach was developed to determine metals an organism assimilates on a given growth medium and identify metalloproteins on a genome- wide scale. Of 343 metal peaks in chromatography fractions, 158 did not match any predicted metalloprotein. Unassigned peaks included metals that *P. furiosus* was known to utilize (cobalt, iron, nickel, tungsten and zinc; 83 peaks) plus metals the organism was not thought to assimilate (lead, manganese, molybdenum, uranium and vanadium; 75 peaks). By shifting the focus from classical protein-based purification to metal-based identification, eight of 158 unexpected metal peaks were purified yielding four novel nickel and molybdenum-containing proteins, whereas four proteins contained sub-stoichiometric amounts of misincorporated lead and uranium. Analyses of two additional microorganisms (*Escherichia coli* and *Sulfolobus solfataricus*) revealed species-specific assimilation of yet more unexpected metals. Metalloproteomes are therefore much more extensive and diverse than previously recognized, and promise to provide key insights for cell biology, microbial growth and toxicity mechanisms (Cvetkovic et al., 2010).

Computational analysis of the large parallel metal and protein dataset (2589 column fractions) yielded predictions of novel metalloproteins in *P. furiosus* (Lancaster et al., 2011). The data are available at http://enigma.bmb.uga.edu/IMPACT. Homologous recombinant production of *P. furiosus* MPs for structural analyses is currently underway. The methodology to identify and predict MPs on a global scale can be adapted and applied to any organism and also provides a road map for the (partial) purification of native forms of novel MPs. We are currently applying this technology to *Desulfovibrio vulgaris*. Large-scale growth (600 L) of the organism and metal determinations using existing chromatography column fractions and correlations with iTRAQ/MS data are currently in progress.

This work was conducted as part of the ENIGMA project and was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Cvetkovic et al. (2010) Microbial metalloproteomes are largely uncharacterized. *Nature* 466, 779
2. Lancaster et al. (2011) A computational framework for proteome-wide pursuit and prediction of metalloproteins using ICP-MS and MS/MS data. *BMC Bioformatics* (in press)

218

Understanding the RNA Landscape in the Microbial World

John Tainer (jat@scripps.edu) and Robert Rambo*

Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: ENIGMA scientists seek to understand in situ microbial activity and community dynamics through detailed assessment of molecular function from proteins to populations. By studying communities with activities of interest to DOE mission we hope to reveal the mechanistic basis for those activities and their support in a changeable and uncertain environment. ENIGMA has 4 main aims

- **Measurement and analysis of environmental activity, composition, structure, and strategies of microbial communities in situ**
- **Use controlled laboratory consortia to identify essential microbial contributions to environmental activities, identify specific and selected interactions, and isolate keystone organisms/processes**

- **Efficiently advance these environmental microorganisms to model organism status and map their molecular functions to community phenotypes and environmental activities.**
- **Development of the LBNL Systems Environmental Microbiology Workbench and Knowledge Framework**

RNA has many diverse roles in microbial biology including direct involvements in transcriptional regulatory elements, RNA modifying enzymes, protein synthesis, intracellular protein trafficking and microbial defense. Therefore, as we aim to establish the microbial biological network for bioengineering, the network must involve a description that includes the role of RNA. Riboswitches such as the S-adenosyl methionine (SAM) riboswitch senses the metabolic environment promoting transcription of an operon in low intracellular SAM conditions. Engineering these microbes for DOE applications requires a systems-biology understanding that must involve the functional annotation of non-coding RNAs (ncRNAs). Our work takes a collaborative approach by leveraging diverse experimental techniques within the native environments of *Desulfovibrio, Halobacterium NRC-1, Pyrococcus furiosus* and *Sulfolobus solfataricus.*

219

Integrated Microbiological Approaches to Characterize Cr(VI)-Reducing Microbial Community at the DOE Hanford 100H Site

Romy Chakraborty* (rchakraborty@lbl.gov), Dominique Joyner, Boris A. Faybishenko, Matthew Fields, Tamas Torok, Gary L. Andersen, and Terry C. Hazen*

Lawrence Berkeley National Laboratory

Project Goals: We have successfully used different approaches in identifying the key microbial players involved in Chromium reduction at the DOE-Hanford 100H site, and are currently developing different strategies to best understand the key metabolic processes mediated by these microbes in the field.

In order to stimulate microbially-mediated reduction of Cr(VI), a poly-lactate compound (HRC) was injected into the chromium-contaminated aquifer at the Hanford (WA) 100H DOE site in 2004. Cr(VI) concentrations rapidly declined to below the detection limit and using high-density DNA 16S rRNA gene microarray (Phylochip), we observed the community to transition through denitrifying, iron-reducing and sulfate-reducing and methanogenic populations. Based on these results, targeted enrichments in defined anaerobic media resulted in the isolation of an iron-reducing *Geobacter metallireducens*-like isolate strain RCH3, a sulfate-reducing *Desulfovibrio vulgaris*-like strain RCH1 and a nitrate-reducing *Pseudomonas stutzeri*-like isolate RCH2 and *Sporotalea* strain 45W among several others. These isolates were capable of reducing Cr(VI) anoxically and the whole-genome sequence data for the first three is now available from JGI. OMNILOG Phenotypic microarray was used to compare isolate RCH1 with the type strain *Desulfovibrio vulgaris Hildenburough (DvH).* The phenotypic microarray allows for high throughput screening of metabolic activity of diverse microorganisms, the panels providing assays for approximately 760 select compounds measuring metabolism of various C, N and P substrates. The high throughput BIOLOG was used for Minimum Inhibitory Concentration (MIC) determinations of environmentally relevant stressors. Further, polyclonal antibodies were raised against the functionally dominant organisms at Hanford including *Methanococcus, Desulfovibrio, Pseudomonas* and *Geobacter* spp and tagged with different fluorescent dyes to enable specific direct enumeration and visualization from environmental samples. Also, streptavidin-coupled paramagnetic beads and biotin labeled antibodies raised against surface antigens of *DvH* were used to capture these type of cells in both bioreactor grown laboratory samples and from Hanford groundwater samples. Field deployable IMS technology may greatly facilitate environmental sampling and bioremediation process monitoring and enable transcriptomics and proteomics/metabolomics-based studies directly on cells collected from the field.

We have successfully used different approaches in identifying the key microbial players involved in Chromium reduction at the Hanford 100H site, and are currently developing different strategies to best understand the key metabolic processes mediated by these microbes in the field.

220

Microbial Community Dynamics from Groundwater and Surrogate Sediments During HRC® Biostimulation for Cr(VI)-Reduction

K.B. De Leon,[1,2,6] B.D. Ramsay,[2,6] D.R. Newcomer,[3] B. Faybishenko,[4,6] T.C. Hazen,[4,6] J. Zhou,[5,6] and M.W. Fields[1,2,6]* (matthew.fields@erc.montana.edu)

[1]Department of Microbiology, [2]Center for Biofilm Engineering, Montana State University; [3]Pacific Northwest National Laboratory; [4]Lawrence Berkeley National Laboratory; [5]University of Oklahoma; and [6]ENIGMA

http://enigma.lbl.gov

Project Goals: Determine bacterial community dynamics during biostimulation for chromate reduction.

The Hanford 100-H site is a chromium-contaminated site that has been designated by the Department of Energy Environmental Management as a field study site for *in situ* chromium reduction. In August 2004, the first injection of hydrogen release compound (HRC®) resulted in an increase of microorganisms and a reduction of soluble chromium(IV) to insoluble chromium(III). Little is understood about the microbial community composition and dynamics during stimulation. The aim of this study is to compare microbial communities of groundwater and soil samples across time and space during a second injection of HRC® via bar-coded pyrosequencing. We have also attempted to validate the pyrosequencing approach to microbial community analysis via the comparison of species richness and diversity estimates to a corresponding clone library for the V4 and V6 regions of SSU rDNA. These results indicate that pyrosequencing data must be thoroughly filtered and that a quality score cutoff is not universal across the SSU rDNA gene likely due to differing proportions of conserved and variable regions.

A second injection occurred November 2008 at the 100-H field site, and geochemical data collected throughout the study showed an overall decrease in nitrate, sulfate, and chromium(IV). Spatial and temporal water and soil samples were collected pre-and post-injection from four wells at the field site. Soil columns constructed from stainless steel mesh were lined with nylon mesh and filled with Hanford soils from the 100-H site. The soil columns were used to represent not only the microbes flowing through the soil via groundwater, but the microbes that require a matrix in order to grow. DNA was extracted from each of the samples and SSU rDNA gene fragments was sequenced via multiplex pyrosequencing. Soil samples differed from the corresponding groundwater (even at the phyla level) and were more diverse. However, although the community composition changed during the biostimulation, the overall community diversity was not altered. Results do not indicate a large shift in dominant organisms in soil from pre- to post- injection, and this may be due to the organisms remaining dominant from the first stimulation. However, a prevalence of core genera and rare genera were observed across 26 samples while urban and rural genera were less abundant. The β-Proteobacteria were more predominant in soil samples while γ-Proteobacteria were more equivalent in both sample types. There was a shift from *Acidovorax* to *Aquaspirillum* from upstream (non-stimulated) to downstream soil both pre- and post-injection. Furthermore, while post-injection soil samples indicate a continuing dominance of *Aquaspirillum*, corresponding water samples indicate *Pseudomonas* as a dominant genus. The greatest changes during stimulation occurred in the populations of mid-dominance either between wells or across time, and these organisms could be important to consider as possible indicator species in future work. Work in progress includes continued phylogenetic structure and composition analyses and characterization of functional diversity via GeoChips.

221

Adaptive Evolution and Physiology of Nascent Microbial Mutualisms

Kristina L. Hillesland,[1]* Birte Meyer,[1]* Nicolas Pinel,[1] Nicholas Elliott,[1] Marcin Joachimiak,[2] Jennifer Kuehl,[2] Adam Deutschbauer,[2] Aifen Zhou,[3] Zhili He,[3] Jizhong Zhou,[3] Dwayne Elias,[4] Terry Hazen,[2] Adam Arkin,[2] and David A. Stahl[1]* (dastahl@u.washington.edu)

[1]University of Washington, Seattle; [2]Lawrence Berkeley National Lab; [3]University of Oklahoma; and [4]Oak Ridge National Lab

http://enigma.lbl.gov
http://stahl.ce.washington.edu/

Project Goals: ENIGMA scientists seek to understand in situ microbial activity and community dynamics through detailed assessment of molecular function from proteins to populations. By studying communities with activities of interest to DOE mission we hope to reveal the mechanistic basis for those activities and their support in a changeable and uncertain environment. In support of those objectives the studies presented here are designed to:

1. **extend understanding of genetic and metabolic networks sustaining and stabilizing natural microbial communities by characterizing different synthetic assemblies of species functioning in a simple two-tier food web of general environmental significance,**
2. **predict end-products of natural selection occurring in a community context,**
3. **develop a bank of genetic variants with known ecological history and evolutionary relationships to enable comprehensive genotype-phenotype map, and**
4. **identify mechanisms of specificity in interactions between species.**

A goal of DOE and ENIGMA is to understand and ultimately predict microbial community responses to environmental change. Key to achieving that goal will be determining the genetic process by which the environment affects population and community characteristics. We studied a simple two-tier microbial food chain composed of *Desulfovibrio* species and hydrogenotrophic methanogens cooperating syntrophically to degrade lactate through the obligate exchange of hydrogen. First, growth and metabolism rates in several pairings of different *Desulfovibrio* and methanogen species were compared and biomass was collected to explore the genetic basis of phenotypic variation among these pairings by microarray analysis. Second, the genetic basis of fitness improvement in an evolving syntrophy between *D. vulgaris* Hildenborough and *M. maripaludis* was explored by genome resequencing and phenotypic comparisons. Evolutionary changes after more than 300 and generations of cooperative growth include significantly increased stability, yield, and growth rate. Illumina sequencing of coculture U9 at 300 generations identified a few molecular differences between both evolved species and their common ancestors. A conserved hypothetical protein (DVU_0799) in *D. vulgaris* with sequence similarities to an outer membrane porin had two mutations (at population frequencies of 100 and 60%). These two mutations change acidic or polar amino acids to non-polar amino acids, and therefore likely affect function. Several other populations that evolved independently but in the same conditions also substituted mutations or in-frame deletions in the same 200 bp region of this gene within the first 300 generations of evolution. Together these results suggest that DVU_0799 has a large beneficial effect on fitness in the evolution environment. *D. vulgaris* clones containing one or both of these mutations could all improve coculture growth, but the magnitude of the effects varied depending on the presence of other mutations. Experiments comparing gene expression provided evidence that a *M. maripaludis* clone from U9 had a differential effect on gene expression in evolved versus ancestral *D. vulgaris*, suggesting that adaptive changes caused specific genetic interactions between evolved *M. maripaludis* and *D. vulgaris*. With continued evolution to 1000 generations, growth rates of evolving cocultures improved substantially. An isolation-independent genome-wide characterization of 12 of these 1000 generation communities using the SOLiD 3 platform identified an average of 10 and 198 mutations at frequencies of ≥70 or ≥25%, respectively, in the 24 species populations. Surprisingly, independent non-/missense mutations were detected frequently in sulfate reduction genes in different *D. vulgaris* lineages. The loss of sulfate reduction capacity by *Desulfovibrio vulgaris*—an organism defined by this characteristic physiology—in multiple lineages has significant implications for better understanding adaptive processes leading to more efficient use of available free energy by microbial communities. Together these data show that some initial evolutionary responses of *D. vulgaris* and *M. maripaludis* to a new, mutualistic environment are repeatable, and that they may affect the interactions between these species. It may thus be feasible to predict some evolutionary responses of species of interest to DOE to environmental change, even when these species are evolving in a community.

222

Parallel Evolution of Transcriptome Structure During Genome Reorganization

Sung Ho Yoon[1]* (syoon@systemsbiology.org), David J. Reiss,[1] J. Christopher Bare,[1] Dan Tenenbaum,[1] Min Pan,[1] Joseph Slagel,[1] Sujung Lim,[2] Murray Hackett,[3] Angeli Lal Menon,[4] Michael W.W. Adams,[4] Adam Barnebey,[6] Steven M. Yannone,[6] John A. Leigh,[2] and **Nitin S. Baliga**[1] (nbaliga@systemsbiology.org)

PI: Nitin S. Baliga[1] (nbaliga@systemsbiology.org)
Co-PIs: John A. Leigh,[2] Murray Hackett,[3] William Whitman,[5] Paul Adams,[7] Adam Arkin,[7] Terry Hazen,[7] Michael W.W. Adams,[4] Greg Hura,[7] Steven M. Yannone,[6] Stephen Holbrook,[7] Gary Siuzdak,[8] and John A. Tainer[7]

[1]Institute for Systems Biology, Seattle, Wash.; [2]Dept. of Microbiology and [3]Dept. of Chemical Engineering, Univ. of Washington, Seattle; [4]Dept. of Biochemistry and Molecular Biology, [5]Univ. of Georgia, Athens; [6]Life Sciences Division, [7]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [8]Scripps Research Institute, La Jolla, Calif.

http://baliga.systemsbiology.net/enigma/
http://baliga.systemsbiology.net/drupal/content/h2-regulation
http://enigma.lbl.gov

Project Goals:

1. **Use transcriptomics, proteomics, and metabolomics to study the systems biology of H_2 metabolism, formate metabolism, nitrogen fixation, and carbon assimilation in *Methanococcus maripaludis*.**
2. **Determine the mechanism of H_2 sensing and transcriptional regulation by H_2.**

Genome streamlining by assembly of genes into operons ("operonization") is instrumental in the continual adaptation of microbes to their environmental niche. However, the random genome reorganization events that drive operonization are also the roots of instability for existing operons. We have determined that there exists a statistically significant trend that correlates degree of operonization in archaea to their phylogenetic lineage. We have further characterized how microbes deal with operon instability by mapping and comparing transcriptome structures of four phylogenetically diverse extremophiles that span the range of operon stabilities observed across archaeal lineages: a photoheterotrophic halophile (*Halobacterium salinarum* NRC-1), a hydrogenotrophic methanogen (*Methanococcus maripaludis* S2), an acidophilic and aerobic thermophile (*Sulfolobus solfataricus* P2), and an anaerobic hyperthermophile (*Pyrococcus furiosus DSM* 3638). We demonstrate how the evolution of transcriptional elements (promoters and terminators) generates new operons, restores the coordinated regulation of translocated, inverted, and newly acquired genes, and introduces completely novel regulation for even some of the most conserved operonic genes such as those encoding subunits of the ribosome. The inverse correlation (r = -0.91) between the proportion of operons with such internally located transcriptional elements and the number of conserved operons in each of the four archaea reveals an unprecedented view into varying stages of the operonization process. Importantly, our integrated analysis has revealed that organism adapted to higher growth temperatures have lower tolerance for genome reorganization events that disrupt operon structures.

This work was supported by U.S. Department of Energy, Award No. DE-FG02-07ER64327 and DG-FG02-08ER64685 (to Nitin S. Baliga); the Office of Science (BER), U.S. Department of Energy, Award No. DE-FG02-08ER64685 (to John A. Leigh); the Office of Science (BES), U.S. Department of Energy, Award No. DE-FG05-95ER20175 (to Michael W. W. Adams). This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231.

223

Systems Biology of Halophiles

Pavana Anur,[1] Justin Ashworth,[1] J. Christopher Bare,[1] Karlyn Beer,[1,4] Manjula Bharadwaj,[1] Aaron Brooks,[1,4] Elijah Christensen,[1] Aimee Desaki,[1] Danielle Durudas,[1] Joseph Horsman,[1] Amardeep Kaur,[1] Fang Yin Lo,[1,4] Bruz Marzolf,[1] Mónica V. Orellana,[1] Min Pan,[1] Wyming L. Pang,[1] Amanda Pease,[1] David J. Reiss,[1] David Rodriguez,[1] Joseph Slagel,[1] Dan Tenenbaum,[1] Serdar Turkarslan,[1] Kenia Whitehead,[1] Elisabeth Wurtman,[1] Sung Ho Yoon,[1] Sujung Lim,[2] **John A. Leigh**,[2] **Murray Hackett**,[3] Angeli Lal Menon,[5] Aleksander Cvetkovic,[6] **Michael W.W. Adams**,[5] Adam Barnebey,[6] Stephen Holbrook,[6] Trent Northen,[6] **John A. Tainer**,[6] **Steven M. Yannone**,[6] **Bonnie Baxter**,[7] **Sunia Trauger**,[8] **Gary Siuzdak**,[8] and **Nitin S. Baliga**[1,2,4]* (nbaliga@systemsbiology.org)

PI: Nitin S. Baliga[1,2,4]* (nbaliga@systemsbiology.org)
Co-PIs: John A. Leigh,[2] Murray Hackett,[3] Michael W.W. Adams,[5] Trent Northen,[6] John A. Tainer,[6] Steven M. Yannone,[6] Bonnie Baxter,[7] Sunia Trauger,[8] and Gary Siuzdak[8]

[1]Institute for Systems Biology, Seattle, Wash.; [2]Dept. of Microbiology, University of Washington, Seattle; [3]Dept. of Chemical Engineering, University of Washington, Seattle; [4]Molecular and Cellular Biology Program, University of Washington, Seattle; [5]Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens; [6]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [7]Westminster College, Salt Lake City, Utah; and [8]The Scripps Research Institute, La Jolla, Calif.

http://baliga.systemsbiology.net/enigma/
http://baliga.systemsbiology.net/drupal/content/enigma
http://enigma.lbl.gov

Project Goals: Developing a cutting edge, multi-scale systems biology framework spanning from single cell to community level interactions to enable predictive modeling of *Halobacterium salinarum* NRC-1 as a model to understand evolution and biomolecular interactions of DOE relevant organisms.

Bioremediation of contaminated environments occurs through the collective metabolism of microbial communities. Strategies to enhance the potential of microbes to detoxify their environment require an understanding of how biological networks and community interactions govern microbial physiology. In salt-enriched environments such as the Hanford site, halophilic extremophiles are primary candidates for detoxification strategies. Here we describe the work of a broad consortium of investigators to develop tools that enable predictive modeling of *Halobacterium salinarum* NRC-1 physiology across biological and evolutionary scales. Starting from a fully annotated genome sequence and abundant systems-level measurements of the transcriptome and proteome, we describe our advances spanning from single-cell modeling of gene regulatory dynamics in subcircuits to interspecies modification of population-level behavior. All of our high-throughput technologies, network modeling algorithms, and software tools have been developed within a framework that is generalizable to other systems. This puts us in a unique position to apply these methods to other species of interest, such as DvH, and to suggest how synthetic modification of microbial physiology and community structure may complement current bioremediation efforts.

This work was supported by U.S. Department of Energy, ENIGMA (Environment and Networks Integrated with Genomes and Molecular Assemblies) Component 3, Award No. DE-FG02-07ER64327.

ENIGMA is a Lawrence Berkeley National Laboratory Scientific Focus Area Program supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics: GTL Foundational Science through contract DE-AC02-05CH11231.

224

Bypassing Signal Activation in the System Wide Mapping of Genes Regulated by Response Regulators

Lara Rajeev, Eric G. Luning, Paramvir S. Dehal, Morgan N. Price, **Adam P. Arkin**, and **Aindrila Mukhopadhyay*** (amukhopadhyay@lbl.gov)

Physical Biosciences Division, Lawrence Berkeley National Laboratory

Project Goals

1. **To map the network of genes that are transcriptionally regulated by two component signal transduction systems in *Desulfovibrio vulgaris* Hildenborough.**
2. **To use experimentally validated binding motifs from *D. vulgaris* to predict functions of two component systems in related sulfate reducing bacteria.**

Two component regulatory systems, comprised of sensor histidine kinases and response regulators, are central to the regulation of stress responses in bacteria. Environmental bacteria especially encode large numbers of putative two component systems and the genes regulated by these systems represent the regulatory networks that impact important natural phenomena such as metal, sulfur, nitrogen and carbon cycling. However, due to lack of knowledge regarding the environmental cues that activate signal transduction, and paucity of methods for high throughput genetic manipulation, these valuable networks remain largely unmapped in most bacteria. We used an *in vitro* array-based DAP-chip (DNA Affinity Purified-chip) method to systematically map the genes regulated by all DNA binding response regulators in the model sulfate reducing bacterium, *Desulfovibrio vulgaris* Hildenborough. Our results from the DAP-chip measurements show at least 200 genes, representing approximately 84 operons, to be regulated by 24 response regulators in *D. vulgaris* Hildenborough, of which only one has characterized orthologs. Our results have allowed us to identify the response regulators involved in the regulation of flagella and pili assembly, lactate utilization, exopolysaccharide synthesis, lipid biosynthesis, and in the responses to low potassium, phosphate starvation and nitrite stresses among others. Gene sets regulated by multiple response regulators forming regulatory networks were also discovered. Finally, using the identified gene sets and orthologs in closely related bacteria, we predicted and experimentally verified binding motifs for 15 of these response regulators. These functional predictions may be applied to related species as well, since the binding site motifs appear conserved for several response regulators.

Funding: This research is funded by ENIGMA, a Scientific Focus Area Program supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Foundational Science through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

225

Development of Metagenomic Technologies for Analyzing Microbial Communities

Qichao Tu,[1] Ye Deng,[1] Zhili He,[1] Hao Yu,[1] Yujia Qin,[1] Aifen Zhou,[1] Jianping Xie,[1] Zhenmei Lu,[1] James Voordeckers,[1] Yongjin Lee,[1] Kai Xue,[1] Joy Van Nostrand,[1] Liyou Wu,[1] Yihuei Jiang,[1] **Terry C. Hazen,**[2] **Paul Adams,**[2] and **Jizhong Zhou**[1,2]* (jzhou@ou.edu)

[1]The University of Oklahoma, Norman, Okla.; and [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://enigma.lbl.gov

Project Goals: ENIGMA scientists seek to understand in situ microbial activity and community dynamics through detailed assessment of molecular function from proteins to populations. By studying communities with activities of interest to DOE mission we hope to reveal the mechanistic basis for those activities and their support in a changeable and uncertain environment. ENIGMA has 4 main aims

- **Measurement and analysis of environmental activity, composition, structure, and strategies of microbial communities in situ**
- **Use controlled laboratory consortia to identify essential microbial contributions to environmental activities, identify specific and selected interactions, and isolate keystone organisms/processes**
- **Efficiently advance these environmental microorganisms to model organism status and map their molecular functions to community phenotypes and environmental activities.**
- **Development of the LBNL Systems Environmental Microbiology Workbench and Knowledge Framework**

Understanding the composition, structure, and interactions of microbial communities in natural environments over time and space is crucial in microbial ecology. We have developed various metagenomics technologies to characterize microbial community structure. First, based on previous GeoChips, we have developed GeoChip 4.0, a more comprehensive GeoChip to facilitate the analysis of microbial communities from a variety of habitats. GeoChip 4.0 contains 120,054 distinct probes, covering 200,393 genes involved in different functional processes important to biogeochemistry, ecology, environmental sciences and human health. Among these, 36,062 probes are specifically designed for the human microbiome, and cover 47,979 genes in 139 functional gene families involved in 19 functional processes. In addition to updating functional gene families from previous versions of GeoChip with the latest NCBI protein repository, 118 new gene families, belonging to bacteriaphage, stress, and virulence, have been added to GeoChip 4.0 to target more microbially mediated functional processes. As a new version, GeoChip 4.0 was developed on the NimbleGen 12x135K platform so that each chip contains 12 arrays, making it possible to hybridize 12 samples under nearly identical conditions at the same time. Computational evaluation of probe specificity indicated that all designed probes were highly specific to their corresponding targets. Experimental evaluation of specificity, sensitivity and quantification using artificial and environmental samples showed GeoChip 4.0 to be a highly specific, sensitive and quantitative tool for microbial community analysis. GeoChip 4.0 has been used to analyze environmental samples from oil spill sites, soil, and human feces and proven to be a rapid and powerful tool in the study of microbial ecology. Also, a random matrix theory-based (RMT) conceptual framework for identifying functional molecular ecological networks was developed with the high throughput functional gene array hybridization data. Our results indicated that RMT is a powerful method to identify functional molecular ecological networks in microbial communities. Elucidating network interactions in microbial communities and their responses to environmental changes is fundamental in research in microbial ecology, systems microbiology, and global change. In addition, amplicon sequencing approaches have been widely used in microbial ecology, but we have found that the reproducibility and quantitative capability are quite low, primarily due to random sampling. Various approaches have been developed to predict and minimize the artifacts associated with random sampling processes. This study will have substantial impacts on microbial ecology because this problem is associated with all current sequencing-based metagenomic studies important to energy production, climate change, environmental management, industry, agriculture, and human health. Our future work will focus on continuously updating the GeoChip with more functional processes and more microorganisms covered, developing data analysis pipelines, and using GeoChip data for molecular ecological network analysis to allow more rapid comprehensive analyses of microbial community composition, structure and functions.

This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231

226

Microfluidic Tools for Single-Cell Genomic Analysis of Environmental Bacteria

Peng Liu,[1] Robert J. Meagher,[1] Yooli K. Light,[1] Suzan Yilmaz,[1] Romy Chakraborty,[2] **Adam P. Arkin**,[3] **Terry C. Hazen**,[2] and **Anup K. Singh**[1]* (aksingh@sandia.gov)

[1]Biotechnology and Bioengineering Department, Sandia National Laboratories, Livermore, Calif.; [2]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: The goal of this project is to develop technologies for genomic analysis of single bacterium found at Hanford site to allow us 1) elucidate the genetic diversity of as-yet uncultivated microorganisms and 2) link function to species, a feat not achievable by current metagenomic techniques.

Current metagenomic techniques (e.g., microarray or 16s rRNA sequencing) relying on pooled nucleic acids from lysed bacteria can independently measure metabolic activity and the species present, but can not link the activity deterministically to the species . We are developing high-throughput tools for studying bacteria one cell at a time, allowing us to unravel the complicated dynamics of population, gene expression, and metabolic function in mixed microbial communities. Our approach includes FISH-based identification of desired species, enrichment by cell sorting, followed by single-cell encapsulation, whole genome amplification and sequencing. Encapsulation of bacteria in pico-liter plugs in particular allows us to scale down conventional (microliter-volume) assays, such as WGA, into much smaller reaction volumes better suited to the size of an individual microbe. By dramatically reducing the reaction volume, the effective concentration of template is increased, reducing amplification artifacts that often arise in single-cell reactions carried out at a conventional scale. These technologies are being used to analyze water samples from Hanford site.

This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231.

227

High-Throughput Mutagenesis Strategies for Non-Model Microorganisms

Adam Deutschbauer[1]* (AMDeutschbauer@lbl.gov), Grant M. Zane,[2] Samuel Fels,[2] Hannah L. Korte,[2] Morgan N. Price,[1] Jennifer Kuehl,[1] Jason Baumohl,[1] **Adam P. Arkin,**[1] and **Judy D. Wall**[2]

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory; and [2]Department of Biochemistry, University of Missouri

https://sites.google.com/a/lbl.gov/enigma/

Project Goals: Our aim it to achieve a deep evidence-based annotation of DOE relevant micoorganisms. To accomplish this we are developing large-scale mutagenesis and phenotyping strategies that are applicable to virtually any microorganism. Our data will be used to fill metabolic holes, uncover functions for hundreds of hypothetical proteins, and to discover novel functional relationships across the entire genome.

Most genes in ENIGMA-relevant bacteria are poorly annotated and many are hypothetical. To address this challenge, it is imperative that flexible, rapid, and inexpensive experimental approaches are developed to assign gene function on a global scale. Here we describe three efforts to annotate gene function in sulfate-reducing and metal-reducing bacteria using high-throughput mutagenesis and phenotyping strategies: (1) directed, markerless genetic modification in *Desulfovibrio vulgaris* Hildenborough (*Dv*H), (2) parallel analysis of transposon mutants using TagModules, and (3) parallel analysis of transposon mutants by HITS (High-Throughput Insertion Tracking by Deep Sequencing). In collaboration with the Chhabra group, we have optimized construct methodologies and can now systematically generate defined *Dv*H mutants using the markerless approach. Our *Dv*H markerless methodology permits the construction of multiple genetic modifications opening the door for the systematic study of genetic interactions in a sulfate-reducing bacterium. Additionally, the markerless system holds great promise for the functional characterization of polymorphisms in evolved lines of *Dv*H (in collaboration with the Zhou and Stahl labs). Despite these advances, however, it is currently not feasible to pursue whole-genome targeted deletion libraries in all ENIGMA microorganisms. Therefore, to augment our markerless deletion approach, we have pursed transposon mutagenesis with the TagModule collection. Each Gateway-cloned TagModule contains two unique 20 bp DNA tags that permit strain pooling and parallel analysis of tag abundance. We combined the TagModules with transposon mutagenesis to create a library of ~50,000 sequence-verified and tagged mutants in *Shewanella oneidensis* MR-1 (~28K mutants), *Desulfovibrio alaskensis* G20 (~15K mutants), and *Dv*H (~7K mutants) which, as archived single strains, serve as a rich resource for all ENIGMA collaborators. Furthermore, the presence of the TagModules permits the pooling and parallel analysis of strain fitness for ~4000 mutants by a highly quantitative, inexpensive assay. We used this pooled assay to probe the fitness of 3355 *S. oneidensis* MR-1 genes (~90% of the nonessential genome) in over 100 diverse growth conditions including different growth substrates, alternative electron acceptors, stresses, survival, and motility. We show that the pooled assay has excellent biological consistency, and relative defects as small as a 2% reduction in growth rate per generation can be assayed reliably. We find that ~70% of genes have a pattern of fitness that is significantly different from random including hundreds of hypothetical genes, and ~37% of genes have a strong enough signal to show strong biological correlations. Using fitness patterns, we were able to propose specific molecular functions for 28 genes or operons that lacked specific annotations or had incorrect annotations including a previously unknown acetylornithine deacetylase. While the TagModule approach described above was accomplished by a single laboratory and can be generally applied to create a large-scale gene-phenotype map in most microorganisms, there is still an upfront investment required to generate the initial mutant strains. To accelerate the analysis of large transposon libraries in additional ENIGMA microorganisms, we are complementing the TagModule approach with the HITS method that does not require the up-front effort to archive single mutants. Preliminary work on developing HITS in *Dv*H is presented.

228

Subcellular Localization of Proteins in the Anaerobic Sulfate Reducer *Desulfovibrio vulgaris* via SNAP-Tag Labeling and Photoconversion

A. Gorur,[1] C.M. Leung,[1] S. Chhabra,[1] T. Juba,[1] A. Tauscher,[1] S. Reveco,[1] J.P. Remis,[1] B. Lam,[1] J.T. Geller,[1] T.C. Hazen,[1] M. Biggin,[1] J.M. Chandonia,[1] K.H. Downing,[1] J.Wall,[1,2] and M. Auer[1]* (mauer@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley Calif.; and [2]University of Missouri, Columbia

Project Goals: Protein localization and expression studies in planktonic cells and biofilms of *Desulfovibrio vulgaris* Hildenborough (DvH) under baseline and environmentally relevant stress conditions

A systems biology understanding of microbes in a planktonic state and in biofilms requires the mapping of spatio-temporal distributions of macromolecules correlated to microbial activity. As part of the DOE-funded ENIGMA program, we study protein expression and localization in the anaerobic soil bacterium *Desulfovibrio vulgaris Hildenborough* (DvH),which plays a prominent role in bioremediation of DOE legacy sites by reducing and therefore immobilizing radionuclide and other toxic heavy metals in plumes therefore preventing these metals from reaching human water supplies. . Our goal is to study protein abundance and localization at the optical level as well as the EM level using cryo-tomography of labeled, photoconverted and vitrified whole-mounts as well as FIB/SEM of resin-embedded samples.

We have chosen the commercially available AGT-tag to label proteins as—unlike GFP and derivatives—it allows labeling under anaerobic conditions. This system is based on a modified O6-alkylguanine-DNA alkyltransferase (AGT) tag that undergoes a dead–end chemical reaction with a modified O6-benzylguanine (BG) derivative. This SNAP label has been conjugated to a large number of fluorophores and other biochemically functional groups allowing flexibility in experimental design.

The tagged strains that are generated using SLIC and Gateway approaches are labeled with commercially available SNAP fluorophores. After extensive optimization we have obtained robust protocols that are virtually background-free and that allow high-throughput imaging and protein expression quantification. Using deconvolution microscopy we have studied ~20 tagged strains and found a significant number to display discrete non-uniform localization patterns. For example, we found ParA, Mot-A and Mot A-1 to localize exclusively to the poles, while others such as Lyt R, FtsH, GlnA, ModA, FlgE and UvrB localize both to the poles and to secondary regions within the cell. Proteins showing a patchy or spotty distribution along the length of the cell include hup-3 and PyrB. As expected the majority of proteins display uniform distribution. We have further developed labeling and photoconversion approaches that allow visualization of protein location in the context of cellular ultrastructure and should allow us to examine its relationship to extracellular metal reduction activity that we discovered to be localized to discrete sites on the outer membrane surface. Only a subset of cells in planktonic state or in biofilms showed metal deposits on the cell surface suggesting that despite seeing the same microenvironment cells differ in their protein inventory and possibly metabolic state.

Acknowledgements/References: This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231.

229

High Throughput Identification of Protein Complexes from *Desulfovibrio vulgaris* by a Tandem Affinity Purification Pipeline

Gareth P. Butland[1]* (GPButland@lbl.gov), **Swapnil R. Chhabra,**[1] Barbara Gold,[1] Nancy L. Liu,[1] Sonia Reveco,[1] Tom R. Juba,[2] **Judy D. Wall,**[2] Bonita R. Lam,[1] Jil T. Geller,[1] **Terry C. Hazen,**[1] Megan Choi,[1] **Mark D. Biggin**,[1] Evelin D. Szakal,[3] Simon Allen,[3] Haichuan Liu,[3] **H. Ewa Witkowska**,[3] and **John-Marc Chandonia**[1]

[1]Lawrence Berkeley National Lab, Berkeley, Calif.; [2]Univ. of Missouri, Columbia; and [3]Univ. of California, San Francisco

Project Goals: see below

ENIGMA's goal is to understand, at a molecular systems level, the bacterial soil communities at DOE sites contaminated with heavy metals or radionuclides. Sulfate reducing bacteria (SRBs) are key members of these communities and can reduce many of the contaminating elements to an insoluble form. Environmental change or human intervention will alter the chemical environment in the subsoil, which in turn affects which species predominate as well as microbial physiology. Therefore, it will be critical to learn how such changes affect SRBs and their interaction with other members of the community and the biogeochemistry. We have chosen *Desulfovibrio vulgaris* to address these questions in molecular detail as it is one of the dominant sulfate reducers found at DOE sites.

Most cellular processes are mediated by multiple proteins interacting with each other in the form of multi-protein complexes and not by individual proteins acting in isolation. In order to accurately model cellular processes in this SRB and how they respond to stress, our goal is to develop a comprehensive knowledgebase of protein complexes and protein-protein interactions in *D. vulgaris* using high throughput tandem affinity purification (TAP). Our approach utilizes the Sequence and Ligation Independent Cloning (SLIC) technique to generate custom suicide constructs in high throughput. Utilizing SLIC, we have achieved success rates for suicide construct generation of greater than 85% and following introduction of constructs into *D. vulgaris* ~80 % of isolates were found to express a TAP-tagged fusion protein by IP-western. Currently, we have generated 687 TAP-tagged *D. vulgaris* strains and for the last six months have been able to generate 50 new TAP-tagged strains per month.

To date, 357 unique *D. vulgaris* strains containing correctly integrated TAP-tagged chromosomal fusions have been subject to TAP analysis and the composition of purified eluates analyzed by mass spectrometry. In 291 of these analyses, the bait was verified to be present by gel-free mass spectroscopy. In these experiments, a total of 5,944 interactions were detected with 1,060 distinct prey proteins. Using curated gold standard datasets, we filtered out ubiquitous proteins and other likely false positives, resulting in a set of 293 high-confidence interactions between 89 baits and 246 preys. 38 interactions have been reciprocally confirmed, using strains in which the original prey protein was tagged and used as bait. Detected high-confidence interactions cover a range of biological processes including energy conservation (Hydrogenase(s), Dissimilatory Sulfite Reductase), protein secretion (YajC-HflCK complex), protein folding (DnaK-DnaJ-DafA complex) and cofactor biosynthesis (Heme and FeS clusters) and include both novel and previously predicted interactions. ENIGMA has also identified a large number of protein-protein interactions in *D. vulgaris* using a tagless approach, and we are integrating the analyses of these data with each other and with other large-scale ENIGMA datasets (e.g., fitness and gene expression data) in order to increase the number of high-confidence interactions. Throughout the project, we have removed many bottlenecks associated with working with *D. vulgaris* and this has enabled us to obtain throughput statistics, data quality and success rates similar to those previously reported for *E. coli*. Our rate of TAP analysis had been limited by the larger culture volumes required for *D. vulgaris* compared to *E. coli*. Improvements in biomass yield, purification processes and mass spectrometry technology have recently enabled a shift to processing 20 strains per week. We are now in a position to conduct a system- wide analysis of all stable protein-protein interactions in *D. vulgaris* and to target how these change in response to stresses typically occurring in the subsoil of contaminated sites for a select set of stress response genes.

This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231.

230

Accurate, High-Throughput Identification of Stable Protein Complexes in *Desulfovibrio vulgaris* using a Tagless Strategy

John-Marc Chandonia[1,3]* (JMChandonia@lbl.gov), Ming Dong,[1] Maxim Shatsky,[1,3] Haichuan Liu,[2] Lee Yang,[1] Terry C. Hazen,[1] Jil T. Geller,[1] Megan Choi,[1] Evelin D. Szakal,[2] Simon Allen,[2] Steven E. Brenner,[1,3] Steven C. Hall,[2] Susan J. Fisher,[1,2] Sunil Kumar,[4] Farris L. Poole,[4] Michael Adams,[4] Jian Jin,[1] H. Ewa Witkowska,[2] Adam P. Arkin,[1,3] and **Mark D. Biggin**[1]

[1]Lawrence Berkeley National Laboratory; [2]University of California, San Francisco; [3]University of California, Berkeley; and [4]University of Georgia

http://enigma.lbl.gov

Project Goals: We describe a novel "tagless" method for identification of stable, soluble protein complexes that is general to all cultureable microbes and does not require genetic manipulation of the organism. Our strategy is based on the premise that the great majority of such com-

plexes will survive intact through a series of orthogonal chromatographic steps, with complex components having correlated elution profiles. We demonstrate the effectiveness of this method in *D. vulgaris*.

Desulfovibrio vulgaris has been selected as a model bacterium for intensive study by ENIGMA because it can reduce heavy metals and radionuclide contaminants present in the soil at many DOE sites, rendering the contaminants insoluble. ENIGMA seeks to model, at a molecular systems level, how this and similar bacteria respond to natural and human induced changes in their environment and how this alters their ability to stabilize contaminants in the soil. A major component our strategy is to develop and use high throughput pipelines to purify and identify protein complexes and to structurally characterize them by EM. Most cellular processes are mediated by multiple proteins interacting with each other in the form of multi-protein complexes and not by individual proteins acting in isolation. Thus, for systems modeling it is critical to characterize protein complexes genome-wide and determine how their composition and structures change with the environment.

We describe a novel “tagless” method for identification of stable, soluble protein complexes that is general to all cultureable microbes and does not require genetic manipulation of the organism. Our strategy is based on the premise that the great majority of such complexes will survive intact through a series of orthogonal chromatographic steps, with complex components having correlated elution profiles. A major challenge is the potential for false positives (FP) caused by co-elution of proteins that are not part of a complex. Approximately 10 g soluble protein from a crude *D. vulgaris* extract have been separated using ammonium sulphate precipitation and a series of three highly parallel chromatographic steps. For the last step, 306 size exclusion columns have been run, yielding 6,859 fractions. The elution profiles of each protein across each of these columns have been measured with the aid of mass spectrometry and iTRAQ reagents (Dong et al., 2008, J Proteome Res. 7:1836-49) leading to the identification of 1,444 proteins (~40% of the proteome). For every region of elution space where two proteins overlap, Pearson correlation coefficients have been calculated between vectors of normalized relative protein amounts estimated using iTRAQ. These data were used to train a random forest classifier to identify true interactions in a manually curated gold standard (GS) set. We compare the discriminating power of these proteomic data to that of other high-throughput data, such as correlation of gene expression profiles. Our method is able to identify 66% of GS interactions present in our proteomic data at a 0% FP rate. Using the same thresholds results in the prediction of 854 novel interactions. Thus, this strategy is effective at identifying a subset of stable inter-protein interactions in a bacterial proteome at high precision.

In addition, we selected 16 complexes identified by the above fractionation strategy with molecular weights 400 - 1,000 kDa and provided them to the ENIGMA single-particle EM group. This resulted in the structures of 7 complexes being solved and showed that there are a surprisingly large number of differences in the quaternary structures of *D. vulgaris* complexes isolated from compared to those of homologous proteins from other microbes (Han et al., 2009, PNAS 106, 16580); see Han et al, poster.

We have also begun to measure the metal content of each of the size exclusion column fractions and compare the results to our iTRAQ quantification of polypeptides to identify metalloproteins; see Menon et al poster.

In the future, we plan to complete our analysis of the full *D. vulgaris* proteome and metalloproteome and study how interactions change under stress conditions that mimic those that commonly occur in contaminated soils. We will also extend our analyses to include other high-throughput datasets produced by ENIGMA, e.g., the protein interactions we have discovered using parallel Tandem Affinity Purification data from *D. vulgaris*; see Butland et al poster.

Funding: ENIGMA is a Lawrence Berkeley National Laboratory Scientific Focus Area Program supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Foundational Science through contract DE-AC02-05CH11231

231

EM Structural Survey of Large Protein Complexes in *Desulfovibrio vulgaris* and EM High Throughput Pipeline Development

Bong-Gyoon Han[1]* (BGHan@lbl.gov), Ming Dong,[1] Maxim Shatsky,[1,2] Pablo Arbelaez,[2] Jitendra Malik,[2] Dieter Typke,[1] Ross Walton,[1] Amos Song,[1] Steven Yannone,[1] Kenneth H. Downing,[1] Mark D. Biggin,[1] and Robert M. Glaeser[1]

[1]Lawrence Berkeley National Laboratory, Berkeley Calif.; and [2]University of California, Berkeley

Project Goals: ENIGMA is conducting a systems level characterization of *Desulfovibrio vulgaris* to understand the role of sulfate reducing bacteria in reducing metals in contaminated DOE sites. As part of that effort, we are establishing methods to structurally characterize multiprotein complexes by single-particle electron microscopy.

Protein samples were purified by a tagless strategy to carry out an unbiased survey of the stable, most abundant multiprotein complexes in *Desulfovibrio vulgaris* Hildenborough (*Dv*H) that are larger than Mr ~400 kD. The quaternary structures for 8 of the 16 complexes purified during this work were determined by single-particle reconstruction of negatively stained specimens (Han et al., 2009, PNAS 106, 16580), a success rate about 10 times greater than that of previous “proteomic” screens. In addition, the subunit compositions and stoichiometries of the remaining complexes were determined by biochemical methods. Our results show that the structures of large protein complexes vary to a surprising extent from one microorganism to another. Except for GroEL and the 70S ribosome, none of the 13 remaining complexes with known orthologs have quaternary structures that are fully conserved. This result indicates that the interaction interfaces within large, macromolecular complexes are much more variable than has generally been appreciated. As a consequence, we suggest that relying solely on quaternary structures for homologous proteins may not be sufficient to properly understand their role in another cell of interest. The diversity of subunit stoichiometries and quaternary structures of multi-protein complexes that has been observed in our experiments with *Dv*H is relevant to understanding how different bacteria optimize the kinetics and performance of their respective biochemical networks.

Conventional single particle EM methods have not previously been able to solve protein structures rapidly enough to handle the sheer volume of protein samples produced by ENIGMA . Therefore, we have reduced the data processing time two fold by automating data collection. To further increase throughput we are implementing automated data analysis, model building from the low tilt angle image pairs and the engineering of new support-film technologies for EM sample preparation. The latter is driven by the need, encountered within this high-throughput project, for technologies that do not require sample-dependent optimization and are more likely to preserve quaternary structure in a conformationally homogeneous state.

232

High Throughput Production and Analysis of Genetically Engineered *Desulfovibrio vulgaris* Strains via Homologous Recombination

PIs participating: Swapnil Chhabra[1]* (srchhabra@lbl.gov), Manfred Auer,[1] Gareth Butland,[1] John Marc Chandonia,[1] Terry Hazen,[1] Judy Wall,[2] Ewa Witkowska,[2] Dwayne Elias,[2] Michael Adams,[3] Matthew Fields,[4] Jan Liphardt,[1] Greg Hura,[1] and David Stahl[5]

[1]Lawrence Berkeley National Laboratory; [2]University of Missouri; [3]University of Georgia; [4]Montana State University; and [5]University of Washington

The primary focus of ENIGMA is to understand, at a basic systems level, bacterial communities in the sub soil of DOE sites contaminated with heavy metals or radionuclides. Sulfate reducing bacteria play a critical role in these communities and can directly reduce many of the contaminants to an insoluble form. Conditions in the sub soil are not static, however. Environmental or human intervention can alter oxygen, nitrate or salinity levels etc, which in turn affects which species predominate and details of their physiology. Understanding how microbes respond to such changes and how sulfate reducers interact with the other members of the community is critical if we are to model how communities cope with such change and interact with the biogeochemistry. We have selected *Desulfovibrio vulgaris,* one of the dominant sulfate reducers found at DOE sites, to understand these processes in molecular detail. For our work, we require many genetically engineered strains in which either affinity tag DNA sequences are introduced into genes, or the activity of the gene is altered by targeted mutation. The ability to modify genomes by making such locus-specific chromosomal alterations in a high-throughput and cost-effective manner has been successfully applied in yeast and *Escherichia coli*, but prior to our work, it was extraordinarily difficult to modify even a few genes in *D. vulgaris* in this way. Indeed, a diverse range of other bacteria of importance to DOE's mission have been similarly difficult to modify.

Therefore, we have developed a method for high-throughput targeted manipulation of genes in *D. vulgaris* that is both inexpensive and flexible due to the use of interchangeable "parts" for making different kinds of chromosomal modifications, including gene deletions, tagged genes for the study of protein-protein interactions and protein localization to name a few. This systematic approach for chromosomal modification can be applied to a wide range of bacteria with minimal need for methodological alteration and relies on the facile construction of suicide vectors through the use of high-throughput methods, including Sequence and Ligation Independent Cloning (SLIC), heretofore used for plasmid-based (rather than chromosomally based) metabolic engineering and heterologous protein expression studies. Our procedures generate tagged genes or marker exchange deletions by double homologous recombination events. For tagged genes, this protocol ensures that a single copy of the gene with the tag is produced from its natural promoter and, in most cases, without polarity. For deletion construction, sequences of the targetted gene are removed from the cell, preventing rearrangements that could restore a functional gene.

Prior to our work, it was only possible to produce a handful of homologous recombination targeted mutations in *D. vulgaris* per year, with the attempts failing for most genes. With our new strategy, we have been able to produce 50 strains per month for the last 6 months and a total of ~762 strains over all. Importantly, over 79% of attempts to modify genes have been successful. Thus, we can now target most genes of importance in various stress responses and are in a position to conduct a full genome-wide analysis. We have engineered 687 strains for Tandem Affinity Purification, which are being successfully used for protein/protein interactions analyses and structural characterization of proteins. A further 75 strains have been constructed for mapping the location of protein complexes within the cell by either deconvolution or single molecule microscopy. In addition, our method will help fill gaps in metabolic pathways and greatly assist in the functional annotation of unknown genes. Our goals for the next few years are to produce strains for the mapping the full *D. vulgaris* interactome and localizome, to assist further annotation of the genome, and to initiate examination of protein complex remodeling in response to environmental stresses.

233

High-Throughput Pipeline for the Purification and Identification of *Desulfovibrio vulgaris* Membrane Protein Complexes

Peter J. Walian[1]* (PJWalian@lbl.gov), Simon Allen,[2] Lucy Zeng,[1] Evelin D. Szakal,[2] Haichuan Liu,[2] Steven C. Hall,[2] Susan J. Fisher,[1,2] Ralph Santos,[1] Bonita Lam,[1] Jil T. Geller,[1] **Terry C. Hazen,[1] John-Marc Chandonia,[1] H. Ewa Witkowska,[2] Mark D. Biggin**,[1] and **Bing K. Jap[1]**

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [2]University of California, San Francisco

http://enigma.lbl.gov

Project Goals: To develop and apply a pipeline for the high-throughput isolation and identification of *Desulfovibrio vulgaris* membrane protein complexes in cultures grown under standard conditions, and to characterize changes in these complexes brought about by environmentally relevant stressors.

The ability of the Gram-negative sulfur-reducing bacterium *Desulfovibrio vulgaris* to reduce heavy metals makes it an ideal candidate for studying how bacteria can influence the biogeochemistry in the subsoil at DOE contaminated sites. Knowledge of this organism's molecular level responses to environmental changes will be essential for accurately modeling its stress response pathways and understanding how it behaves under a range of chemical environments. To obtain this knowledge we have developed a high-throughput pipeline for the isolation and identification of untagged membrane protein complexes. Membrane proteins are especially of interest, as they serve at the interface of cell-cell communications and coordinate interactions with the extracellular environment. It is widely recognized, however, that these proteins are particularly challenging to purify and characterize, and that the use of an inappropriate detergent or detergent-to-protein ratio, for example, can lead to the disruption of complexes or the formation of non-biologically relevant aggregates. The relatively low yield of *D. vulgaris* cells per liter of culture (about one-tenth that of *E. coli*) presents an additional challenge. To address these challenges, we have developed a unique processing pipeline that features mild, but effective, detergent solubilization, liquid chromatography and native electrophoresis methods. Large-scale cultures up to 100 liters in size have been processed. This "tagless" strategy can provide uniquely global views of changes in membrane protein populations from cultures grown under a variety of conditions. Additionally, a distinct advantage of the tagless approach is that it does not require genetic manipulation which can invoke a steep and time-consuming learning curve when tackling organisms where there is minimal previous genetic experience. Thus, our method should be general to a wide array of microbes of interest to the DOE but for which to date facile genetic manipulation is not possible.

Our tagless study of the membrane protein complexes of the *D. vulgaris* outer-membrane is complete and we are at an advanced stage of data collection for the inner-membrane component with over 1000 samples being analyzed by mass spectrometry per growth condition. Outer-membrane proteins provide the front line of defense for Gram-negative bacteria such as *D. vulgaris* and are expected to readily yield a variety of membrane protein changes in response to environmental stressors. We have identified 70 outer-membrane proteins derived from cells grown to mid-log phase under standard conditions representing a coverage of 82% of those predicted to be encoded by the genome. Of these, at least 50 appear to be in some form of complex (homo- or heteromeric). This list will serve as our reference dataset with which we can assess the nature of a cell's membrane protein-based response to a variety of stresses. Processing of outer-membrane proteins from stressed cultures (to include growth to stationary phase, and growth under elevated levels of nitrate or NaCl) is now underway. Exciting preliminary results provide a global view of a large number of stress-induced changes in outer-membrane proteins and demonstrate the potential of this approach. Taken together, these results indicate that this tagless technique will be an effective tool for revealing changes in *D. vulgaris* membrane proteins arising from environmental stress. We propose to extend our studies to additional stresses that have also been examined by other system wide methods within ENIGMA, and to conduct a comprehensive characterization of the hundreds of inner-membrane proteins encoded by *D. vulgaris*.

234

ENIGMA-MS: Using Stable Isotopes and Novel Metabolomic Technologies to Advance Our Understanding of Microbial Metabolism

Richard Baran[1]* (RBaran@lbl.gov), Peter Benke,[1] Edward Baidoo,[1] **Jay Keasling,[1] H. Gary Siuzdak,[2]** and **Trent Northen**[1]

[1]Lawrence Berkeley National Lab, Berkeley Calif.; and [2]The Scripps Center for Mass Spectrometry, La Jolla, Calif.

http://enigma.lbl.gov

Project Goals: A microorganism-based approach to address DOE mission goals in remediation, carbon sequestration and energy production will require quantitative understanding of biological complexity at multiple scales — from molecular networks of individual species to the dynamic inter-species interactions within the communities in which they reside. The broad goals of ENIGMA are to understand biological complexity by discovering and predictively modeling interactions within microbial and community processes that drive complex geochemistries in key environments. In doing so we expect to define biological principles governing selection of microbial community function and composition in given environments.

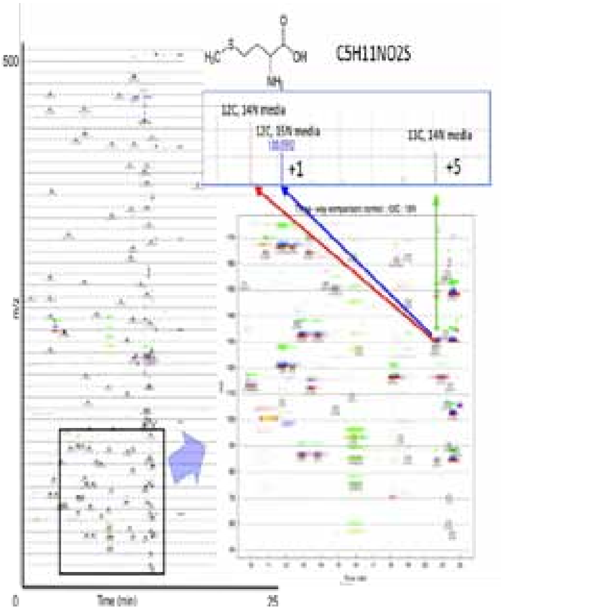


Figure 1. Stable Isotope Labeling for metabolite profiling

Metabolite profiling using mass spectrometry provides an attractive approach for the interrogation of cellular metabolic capabilities. Untargeted metabolite profiling using ElectroSpray Ionization (ESI) has the potential to identify numerous novel metabolites, however, de novo identification of metabolites from spectral features remains a challenge given the large number of experimental artifacts. The ENIGMA MS group has developed and reported an integrated workflow for metabolite identification using uniform stable isotope labeling. Metabolite profiling of cell and growth media extracts of unlabeled control, ^{15}N, and ^{13}C-labeled cultures of the non-filamentous cyanobacterium, *Synechococcus sp. PCC 7002* was performed using normal phase liquid chromatography coupled to mass spectrometry (LC-MS). Visualization of three-way comparisons of raw datasets highlighted characteristic labeling patterns for metabolites of biological origin allowing exhaustive identification of corresponding spectral features (Fig. 1). Additionally, unambiguous assignment of empirical formulas was greatly facilitated by the use of stable isotope labeling. Empirical formulas of metabolites responsible for redundant spectral features were determined and fragmentation (MS/MS) spectra for these metabolites were collected. Analysis of acquired MS/MS spectra against spectral database records led to the identification of a number of metabolites absent not only from the reconstructed draft metabolic network of *Synechococcus sp. PCC 7002*, but not included in databases of metabolism.

This work has recently been extended with systematic screening of consumed or excreted metabolites using metabolite profiling of growth media from microbial cultures (referred to as metabolic footprinting). We performed a systematic evaluation of exchange of metabolites between a *Synechococcus sp. PCC 7002* and different growth media using metabolomics. It was found that 102 out of 202 detected metabolites were exchanged significantly. This metabolic footprinting approach is being extended to study interactions between different organisms. In addition, the presence of membrane transport activities for specific metabolites was highlighted and can enable the search for corresponding transport proteins.

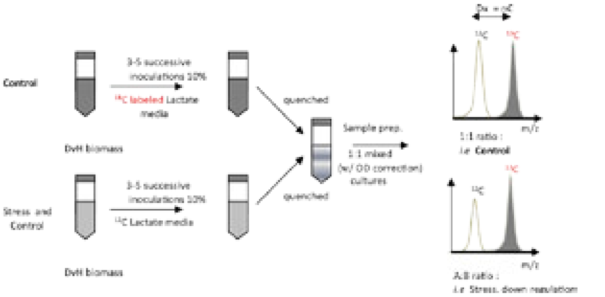


Figure 2. Stable Isotope Labeling based quantification

In addition to facilitating identification of compounds and their exchange with the environment, growth on stable istopic media is used to quantify microbial responses to environmental stresses. Since the quantification of a large number of metabolites is impractical and time consuming, our laboratory has designed an isotope dilution mass spectrometry (IDMS) strategy (Fig. 2) that improves upon precision (as well as on identification, as previously described) and hence simplifies quantification of microbial responses to stress. This strategy involves the mixing of *D. vulgaris* cultures grown side-by-side on the unlabeled (stress) and ^{13}C uniformly labeled (control) form of the same carbon source (lactate). Consequently, the comparison of the ^{12}C and ^{13}C fully-labeled metabolites is representative of stressed versus control biomass. The mixing of the quenched cultures prior further sample preparation enhances the precision of the measurement, which is imperative to quantification with high degree of confidence. Currently, we have achieved ~80% labeling efficiency after three consecutive inoculums (e.g. initiate the feeding of labeled lactate after the cells have entered a period of starvation). However, it is not necessary to achieve 100% ^{13}C conversion, if reproducible labeling efficiencies for all metabolite of interest can be achieved.

235

ENIGMA-MS: Protein Identification, Quantitation and Structural Characterization

Simon Allen,[1] Megan Choi,[2] Haichuan Liu,[1] Christopher Petzold,[2] Max Shatsky,[2] Evelin D. Szakal,[1] Sunia A. Trauger,[3] Lee Yang,[2] Ming Dong,[2] Peter Walian,[2] Steve S. Yanonne,[2] Lucy Zeng,[2] Rich Niles,[1] Susan J. Fisher,[1] Steven C. Hall,[1] **John-Marc Chandonia**,[2] **Jian Jin**,[2] **Gareth Butland**,[2] **Bing Jap**,[2] Paul Benton,[2] Benjamin Bowen,[2] Farris Poole,[2] **John A. Tainer**,[2,3] **Mark Biggin**,[2] **Trent Northen**,[2] **Michael Adams**,[4] Aindrila Mukhopadhyay,[2] **Gary Siuzdak**,[2,3] and **H. Ewa Witkowska**[1]* (witkowsk@cgl.ucsf.edu)

[1]University of California, San Francisco; [2]Lawrence Berkeley National Lab; [3]The Scripps Research Institute; and [4]University of Georgia

Project Goals: The overarching goal of the MS Proteomics component of ENIGMA is to develop and provide MS tools for a comprehensive characterization of proteomes and interactormes of bacteria, with a focus on addressing key challenges that are presented by the bioremediation needs at metal-contaminated sites, e.g., Hanford. To this end, high throughput pipelines for fractionation and identification of protein complexes were introduced and successfully executed for DvH, P. furiosus and S. solfataricus. In addition, the MS-based assays aimed at detailed characterization of protein interactions with protein and/or non-proteinaceous partners are being developed, e.g., metalloproteomics, identification of proteins interacting with specific ligands and analysis of intact protein complexes, the latter to establish their stoichiometries and architecture. Ultimately, in collaboration with other components of ENIGMA, we plan to develop multi-omic platforms capable of integrated analysis of different factors.

Understanding protein – function relationships and protein interactions with other cell components is critical to ENIGMA's experimental goals. Robust mass spectrometry (MS) platforms provide the primary analytical techniques for identification, quantitation and characterization of bacterial proteomes. Our group has implemented novel experimental workflows, introduced automated high-throughput routines to enhance MS data acquisition, developed new technologies to increase sensitivity and broaden applicability of MS techniques to protein-associated species, and built computational and bioinformatics tools for data analysis and interpretation. Protein identification is performed routinely using gel- and solution-based liquid chromatography (LC) MS workflows. A number of specialized technologies that apply the power of MS for detection and identification of protein-protein (tagless[1] and qTagless[2]), protein-metal (metalloproteomics)[3] and protein-ligand[4] interactions were developed and applied to various bacterial species. The latter technologies represent integration of metabolomics and proteomics strategies.

Two major workflows are currently available for protein identification (ID): gel- and solution-based LC MS. In the gel-based workflow, final protein separation is performed using electrophoresis (SDS PAGE), visible bands (stained with Coomassie Blue or silver) are excised, proteins are in-gel digested with trypsin in a 96-well format by a robotic platform, and resulting peptides are analyzed via 1D (low pH reversed phase) nanoLC ESI MS, primarily using a Thermo LTQ mass spectrometer operated in a data-dependent mode. Batch data analysis is performed using the Mascot search engine. A variant of the gel-based LC workflow that employs gel-free electrophoresis for protein separation was developed at LBNL: it offers the ability to further automate the process of sample preparation for MS[5].

The gel-based LC workflow is routinely used in a number of ENIGMA projects. Specifically, it has been applied in conjunction with metalloproteomic studies to map metal-binding protein complexes in bacterial cell lysates of *Pyrococcus furiosus*. Diethylaminoethyl anion exchange (DEAE) column chromatography-fractionated proteins were analyzed to derive protein IDs and in parallel, with inductively coupled plasma MS to detect and quantitate metal ions present in protein complex-containing fractions. Using this platform, over 7000 samples have been analyzed to identify protein complexes and unique metalloproteins. Gel-based LC MS also serves as a primary workflow for identification of outer- and inner-membrane protein complexes in *Desulfovibrio vulgaris* Hildenborough (*Dv*H). Fractionation of inner and outer membrane complexes in *Dv*H employs a combination of chromatographic and electrophoretic (Blue Gel) steps, with a final separation of protein complex components using 1D SDS PAGE that is followed by MS protein identification using either MALDI or ESI-based methods. To date, ~4000 membrane samples were analyzed. The first survey of protein complexes present in the outer membrane of *Dv*H is currently being prepared for publication. Furthermore, the gel-based LC workflow also serves as an adjunct technique in analysis of pulldowns derived from tandem affinity purification (TAP) experiments.

For solution-based LC MS workflows, tryptic digestion is performed using either classical solution digestion protocols or on a PVDF membrane in a 96-well format. Peptides are analyzed either by 1D or 2D LC MS, the latter utilizing cation exchange or reversed phase separation at basic pH as the first dimension. Solution-based 1D LC MS is also used for analysis of protein mixtures derived from the TAP (>380 baits analyzed so far) and qTagless strategy workflows for the identification of the soluble *Dv*H protein complex components. In addition, 2D LC methods are used for an exhaustive proteomics surveys of soluble and membrane compartments of a bacterial cell.

Different approaches to a tagless, *i.e.*, non-targeted, analysis of protein complexes in bacteria have been utilized and customized for soluble and membrane compartments in different types of bacteria. The protein complex mapping of *P. furiosus* and *S. solfataricus* employed a multidimensional separation of biomass under native conditions and monitoring the protein content of chromatographic fractions by a combination of 1D SDS PAGE and LC MS utilizing the Thermo LTQ mass spectrometer. To analyze soluble protein complexes in *Dv*H, a quantitative qTagless strategy was developed using MS tools to track protein elution through the final steps of a multi-dimensional protein separation space. To this end, dense sampling of protein fractions collected at the size exclusion step of protein fractionation was performed and relative concentrations of each polypeptide across the separation column were measured with the aid of iTRAQ reagents[2]. The derived polypeptide elution profiles were subjected to computational analysis to assign probabilities for "true" protein complex components vs. "opportunistic" coeluters. Currently, all MS analyses for the qTagless strategy utilize a LC MALDI MS/MS workflow (AB 4800 TOF/TOF mass spectrometer). To enhance throughput of the qTagless strategy, a miniaturized protocol for protein digestion and peptide labeling with iTRAQ reagents in the 96-well PVDF membrane plate format was introduced. In addition, automated iterative MS/MS acquisition (IMMA) software was developed to increase the efficiency of protein identification in LC MALDI MS/MS workflows.

The current proteomics workflows that focus on elucidation of changes in the repertoire and in the level of protein expression are not well suited to tackle the subtle, often non-stoichiometric alterations in protein structure, *e.g.*, post-translational modifications (PTMs). At the same time, genomic and transcriptional analyses provide little help in discerning their presence and localization. While a number of approaches will be necessary to provide truly comprehensive protein characterization, we propose to focus on the following areas: (1) targeted and quantitative analysis of phosphorylation and glycosylation as likely drivers of protein function, (2) analysis of intact proteins and protein assemblages as an entry to protein population studies, and (3) integration with stable isotope probes to understand the dynamics of protein expression and link with metabolic capabilities. To this end, the results of pilot analyses of intact protein complexes by native MS are very encouraging. In a published study of PTMs, trimethylation was observed in a number of proteins engaged in sulfate reduction in *Dv*H[6]. Accordingly, there is a need for new technologies that will enable multifaceted characterization of proteins of interest and ultimately, entire proteomes to capture microheterogeneity of structures that can be linked to function. Development of MS tools for targeted PTM discovery, analysis of intact proteins and their interactions with protein and non-protein (*e.g.*, metals, ligands) partners will be prioritized for development by the ENIGMA-MS group.

References

1. Menon et al., 2009. *Mol. Cell. Proteomics* 2009, 8:735-751.
2. Dong et al., 2008. *J Proteome Res.* 7:1836-49.
3. Cvetkovic et al., 2010. *A Nature* 466:779-82.
4. Kalisiak et al., 2009. *J. Am. Chem. Soc.* 131:378-386.
5. Choi et al., 2010. *Electrophoresis* 31: 440-7.
6. Gaucher et al., 2008. *J Proteome Res.* 7: 2320-31.

236

Deconvoluting Signal From Noise: Deciphering Biological Functions and Interactions

Ben Bowen[1]* (bpbowen@lbl.gov), Marcin Joachimiak,[1] David J. Reiss,[3] Morgan Price,[1] John-Marc Chandonia,[1] Paramvir Dehal,[1] Gary Siuzdak,[4] Trent Northen,[1] Adam Arkin,[1] and Nitin Baliga[3]

[1]Lawrence Berkeley National Lab; [2]University of California - Berkeley; [3]Institute for Systems Biology; and [4]The Scripps Research Institute

Project Goals: Overview of the computational tools within the ENIGMA SFA

ENIGMA is at the forefront of systems biology of microbes and their communities. In systems biology, computation has a crucial role in processing large amounts of data to construct a quantitative and predictive understanding of biological function at multiple scales. Initially, our algorithms analyze raw data from a diverse array of high-throughput technologies such as mass spectrometry, sequencing, and high density microarrays to yield quantitative measurements of sequence variations, transcripts, proteins, and metabolites. At an intermediate level, our algorithms integrate these data to find statistically significant patterns over multiple dimensions of environmental space and time. These patterns reveal biodiversity in a community, genome organization in a microbe, transcriptome structure and regulation, protein-protein, regulons, and novel metabolic capabilities. One level up, we are inferring organizational principles that relate the behavior of organisms in a community, and the functioning of regulons and protein complexes within metabolic and regulatory networks. We illustrate examples of efforts within ENIGMA that span this continuum of algorithm development across multiple scales of systems biology:

TIER 1: PROCESSING AND ANALYSIS OF RAW DATA. The ENIGMA project has generated gene expression, gene fitness, proteomic, metabolomic, and protein-protein interaction data among others. A variety of approaches including associative biclustering have found relationships in different biological contexts including: transcription regulatory networks, protein-protein interaction networks, and metabolic pathways. These analysis have led to numerous discoveries, but they are all rooted in the correct handling of complex datasets that present technical and scientific challenges to process.

TIER 2: STATISTICAL PATTERN IDENTIFICATION OF SPATIO-TEMPORAL PHENOMENA. With the ongoing deluge of functional genomics data it has become advantageous to: a) simultaneously query multiple data types, b) jointly determine confidence across data layers, and c) systematically form hypothesis from multiple types of evidence. These challenges are met with approaches involving associative biclustering methods that identify a variety of types of coherent patterns in combined functional genomics data. The method searches for associations using statistical criteria functions and estimates confidence across multiple data types. We have shown that on a model synthetic gene expression dataset our method outperforms other methods designed to identify transcription modules in gene expression data alone. The method is placed in a computational framework, which allows rapid customization and deployment for new data sets and data types. For example, results of searches incorporating data on transcription, such as gene expression microarrays, provide direct information on putative regulons. We are currently analyzing our results of associative data patterns from a large yeast data compendium. We have also developed a graphical viewer to allow interaction with the results of the associative biclustering searches and the associated data types, and we are working on applying this to a massive collection of metabolite mass spectra measurements.

TIER 3: SYSTEMS LEVEL ANALYSIS OF METABOLISM AND REGULATION. Gene regulatory networks (GRNs) spatiotemporally regulate cellular physiology to optimize resource utilization, maintain integrity of genetic information, and contribute towards competitive fitness of the organism under changing environmental conditions. These networks are dynamically modulated with changing environments, and the underlying mechanisms behind these changes may be learned by integrating a wide variety of experimental and genomic data. We have successfully inferred causal and predictive models for these networks in *Halobacterium salinarum* NRC-1, and are currently applying these methods to other DOE-relevant organisms.

237

An ENIGMA Analysis Platform for Metabolomics: Towards Identification of Metabolites from Untargeted Experiments

Gary Siuzdak* (siuzdak@scripps.edu)

[4]The Scripps Research Institute

Project Goals: Develop metabolomics based microbial tools such as XCMS metaXCMS, chemical approaches, and to facilitate identification we have developed a novel database, METLIN. The METLIN Metabolite Database is a repository of metabolite information as well as tandem mass spectrometry data. METLIN is a metabolite database for metabolomics containing over 42,000 structures, it also represents a data management system designed to assist in a broad array of metabolite research and metabolite identification by providing public access to its repository of current and comprehensive MS/MS metabolite data.

Mass spectrometry-based untargeted metabolomics requires data preprocessing approaches to correlate specific metabolites to their biological origin. XCMS is an LC/MS-based data analysis approach incorporating novel nonlinear retention time alignment, feature detection, and feature matching. Without using internal standards, the method dynamically identifies hundreds of endogenous metabolites for use as standards, calculating a nonlinear retention time correction profile for each sample. Following retention time correction, the relative metabolite ion intensities are directly compared to identify changes in specific endogenous metabolites.

XCMS, however, often results in the observation of hundreds to thousands of features that are differentially regulated between sample classes. A major challenge in interpreting the data is distinguishing metabolites that are causally associated with the phenotype of interest from those that are unrelated but altered in downstream pathways as an effect. To facilitate this distinction, here we describe new software called metaXCMS for performing second-order ("meta") analysis of untargeted metabolomics data from multiple sample groups representing different models of the same pheno- type. While the current version of XCMS was designed for the direct comparison of two sample groups, metaXCMS enables meta-analysis of an unlimited number of sample classes to facilitate prioritization of the data and increase the probability of identifying metabolites causally related to the phenotype of interest. metaXCMS is used to import XCMS results that are subsequently filtered, realigned, and ultimately compared to identify shared metabolites that are up- or down-regulated across all sample groups. We demonstrate the software's utility with halobacterium mutants. metaXCMS is freely available at http://metlin.scripps.edu/metaxcms/.

To further facilitate the assignment of unknown mass spectral features, we have also demonstrated that profiling can be performed on cultures uniformly labeled with stable isotopes of nitrogen (^{15}N) or carbon (^{13}C). This makes it possible to accurately count the number of carbon and nitrogen atoms in each molecule, providing a robust means for reducing the degeneracy of chemical space and thus obtaining unique chemical formulae for features measured in untargeted metabolomics having a mass greater than 500 Da, with relative errors in measured isotopic peak intensity greater than 10%, and without the use of a chemical formula generator dependent on heuristic filtering. These chemical formula can serve as indicators for the presence of particular metabolic pathways.

In conjunction with XCMS and metaXCMS, and these chemical approaches, to facilitate identification we have developed a novel database, METLIN. The METLIN Metabolite Database is a repository of metabolite information as well as tandem mass spectrometry data. METLIN is a metabolite database for metabolomics containing over 42,000 structures, it also represents a data management system designed to assist in a broad array of metabolite research and metabolite identification by providing public access to its repository of current and comprehensive MS/MS metabolite data. An annotated list of known metabolites and their mass, chemical formula, and structure are available on the METLIN website. Each metabolite is linked to outside resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) for further reference and inquiry. MS/MS data is also available on many of the metabolites. The list is expanding continuously as more metabolite information is being deposited and discovered.

238

ENIGMA Knowledgebase: MicrobesOnline, Gaggle and RegTransBase

Paramvir S. Dehal[1]* (psdehal@lbl.gov), J. Chris Bare,[2] Pavel S. Novichkov,[1] John-Marc Chandonia,[1] David Reiss,[2] Morgan N. Price,[1] Keith Keller,[1] Jason Baumohl,[1] Marcin P. Joachimiak,[1] Inna L. Dubchak,[1] Adam P. Arkin,[1,3] and Nitin S. Baliga[2]

[1]Lawrence Berkeley National Laboratory; [2]Institute of Systems Biology; and [3]University of California, Berkeley

Project Goals: ENIGMA scientists seek to understand in situ microbial activity and community dynamics through detailed assessment of molecular function from proteins to populations. By studying communities with activities of interest to DOE mission we hope to reveal the mechanistic basis for those activities and their support in a changeable and uncertain environment. ENIGMA has 4 main aims

- **Measurement and analysis of environmental activity, composition, structure, and strategies of microbial communities in situ**
- **Use controlled laboratory consortia to identify essential microbial contributions to environmental activities, identify specific and selected interactions, and isolate keystone organisms/processes**
- **Efficiently advance these environmental microorganisms to model organism status and map their molecular functions to community phenotypes and environmental activities.**
- **Development of the LBNL Systems Environmental Microbiology Workbench and Knowledge Framework**

The ENIGMA Knowledgebase integrates the diverse data sets generated by the project with the central goal of enabling development of computational algorithms to predict gene, microbe and microbial community function. To this end, we maintain systems for low level per experiment data capture, cross experiment data integration and data analysis. Because of the importance of relating experimental data and meta-data, we have created LIMS and relational databases for data repositories for ENGIMA experiments. This underlying data, together with relevant publicly available data sets, is then integrated into MicrobesOnline, Gaggle and RegTransBase to enable the computation of predictive models of metabolism, gene regulation and cell response to environmental stimuli.

MicrobesOnline: The MicrobesOnline database (http://www.microbesonline.org) currently holds over 3000 microbial genomes and is updated semi-annually, providing an important comparative and functional genomics resource to the community. MicrobesOnline continues to provide an interface for genome annotation, which like all the tools reported here, is freely available to the scientific community. MicrobesOnline allows the user to quickly access functional genomics data in comparative and evolutionary framework by providing gene homology, domains, phylogenetic trees, operon and regulons predictions together with functional data such as protein-protein interaction, microarray expression data and phenotype/genotype associations. We have developed methods, FastBLAST, FastHMM and FastTree, to enable us to deal with the many millions of gene sequences generated from metagenomics. These tools allow MicrobesOnline to provide the only comparative metagenomic data browser which features a tree browser for every gene family.

RegTransBase: We have built tools and resources for studying regulation in bacteria and archaea using comparative genomics approach. In addition to working on a high quality semi-manual regulon inference in a wide range of species we are building several on-line resources covering different aspects of regulation. RegTransBase, a database of regulatory interactions from literature collected by a group of experts, currently includes 5,100 annotated articles describing twelve thousand experiments. RegPrecise describes manually curated computational predictions of regulons in bacterial genomes done by comparative genomics. RegPredict is a set of highly integrated web tools for fast and accurate inference of regulons. All regulation-related resources are based on the MicrobesOnline data.

Gaggle: The ability to seamlessly interoperate across analysis tools, data sets and data types developed by different scientists across the world has long been a limitation for biologists. Gaggle is a framework for interoperability between bioinformatics software tools which solves this problem. Gaggle enables message passing between data sources, analysis software and visualization tools, including Cytoscape, MultiExperiment Viewer and R, in addition to web resources such as KEGG and STRING. Gaggle and the Firefox web browser extension Firegoose, allows the user to treat independently developed programs as a larger, coupled suite of tools for exploratory data analysis. In addition to visualization in externally developed tools such as MeV and Cytoscape, the Gaggle Genome Browser was developed to visualize any experimental data along the genome coordinates. Visualization of data such as microarray, RNA-seq, protein mass spectrometry and ChIP-chip/seq in the context of the genome can reveal the molecular mechanisms that regulate transcription.

239

Microbial Deconstruction Proteomics: Mapping Protein Subcellular Localization in the Extremophiles *Pyrococcus furisiosus*, *Halobacterium salinarum*, and *Sulfolobus solfataricus*

Robert Rambo,[1] Adam Barnebey,[1] Michael W.W. Adams,[3] Christopher Bare,[5] Nitin S. Baliga,[5] Trent Northen,[1] Ben Bowen,[1] Sunia Trauger,[4] Gary Siuzdak,[4] John A. Tainer,[1,6] and **Steven M. Yannone**[1]

[1]Dept. of Molecular Biology, Lawrence Berkeley National Lab; [2]Center for Life in Extreme Environments, Portland State University; [3]Department of Biochemistry and Molecular Biology, University of Georgia; [4]Center for Mass Spectrometry, The Scripps Research Institute; [5]Institute for Systems Biology; and [6]Department of Biochemistry and Molecular Biology, The Scripps Research Institute

Project Goals: This project started as part of a larger foundational science program (MAGGIE) to develop tools and technologies needed to manipulate non-model organisms and microbial communities to address DOE mission goals. Given that a large proportion of genes and proteins from organisms of interest are poorly characterized, and that all novel enzymes are uncharacterized by definition, we set out to develop universally applicable and practical technologies for mapping protein localization and protein-assemblies within any given microbe. Our goals include; 1) developing universal microbial deconstruction and fractionation processes that allow proteome-wide analysis retaining abundance, assembly, and localization information for each protein, 2) to develop a user-friendly interface for these complex and large data sets that are useful to both informaticists and bench scientists, 3) develop molecular biology for our model system (*Sulfolobus solfataricus*) to validate novel findings with our methods.

The speed and efficiency of microbial genome sequencing has far outpaced our ability to assign functions to novel genes. To fully exploit the diversity of chemistries evolved within microbial life, and to understand interactions within communities, new types of informative datasets are needed to annotate the growing number of genes with unknown function. Subcellular localization and assembly into larger complexes are informative factors in predicting or determining protein functionality. Extremophilic enzymes and protein complexes are exceptionally stable and arguably the most tractable model system for proteome-wide isolation of macromolecular assemblies.

Here we have chosen the extremophilic microbes *Sulfolobus, Pyrococcus* and *Halobacterium* as model systems to develop a universally applicable and practical technology for mapping protein localization and protein-assemblies within any given microbe. We have applied robust biochemical fractionation techniques coupled to HTP MS/MS proteomics to assign cellular locality and physical characteristics to all proteomic identifications. Density, mass, and cellular locality were exploited to fractionate microbial proteomes for each of these highly divergent extremophiles. Constant buffer conditions matched to the widely varied intracellular condition of each organism were used to both stabilize assemblies, and establish the universal applicability of our approaches. Microbial biomass was partitioned into four primary fractions, 1) extracellular, 2) membrane, and two intracellular fractions 3), high-mass particles, and 4) low-mass particles. With this approach we observe 60% of the predicted *Sulfolobus* proteome (1783 proteins) and 305 proteins partition with >95% confidence of being exclusively in one cellular partition. Not surprisingly, the majority (184) of these reside exclusively in the membrane fraction with 30% of these being hypothetical proteins. The intracellular high-mass partition contained intact thermosome and ribosome that were characterized by small-angle x-ray scattering and EM. The small-mass partition was by far the most complex and degenerate, and contained only 22 proteins that were not observed elsewhere in the cell. These highly complex proteomic data sets are presented for simple and intuitive visual inspection using a genome browser developed within ENIGMA. The methods and technologies developed here are applicable to any organism or community of interest to DOE and provide novel proteome-wide datasets for assigning protein locality, function, and assembly states within microbes.

Because our localization proteomic data sets are novel, they require a means for validation. To address this, we have developed a new high-throughput protein expression system for *Sulfolobus*. We have built on viral based vectors to develop a PCR-based gateway-cloning vector. We are implementing our new molecular biology to validate assemblies inferred from the cellular deconstruction analyses and to validate localization and assembly states of hypothetical and annotated proteins.

Sulfolobus solfataricus is a single cell organism that thrives at 80°C in highly acidic volcanic springs (pH=2-3). There are very few life forms able to compete in this extreme environment, which leads to an exceptionally simple microbial community, including only viral pathogens and fewer than twenty putatively identified organisms. Such a simple community can provide an excellent platform to test hypotheses about co-evolutionary adaptations and community interactions from more complex and less malleable communities. Together, the simple nature of solfataric spring communities, our novel molecular biology, and our deconstruction data sets make *Sulfolobus* a particularly useful model system for testing co-evolution and community interaction hypotheses from more complex systems.

239A‡

submitted post-press

Genome-Scale Phylogenetic Function Annotation of Large and Diverse Protein Families

Barbara E. Engelhardt,[1,6] Michael I. Jordan,[1,2] Susanna Repo,[3] Gaurav Pandey,[3] Ameet Talwalkar,[1] Jeffrey Yunes,[4] **John-Marc Chandonia**,[3,5*] and Steven E. Brenner[3,5] (brenner@compbio.berkeley.edu)

[1]EECS Department, University of California, Berkeley; [2]Department of Statistics, University of California, Berkeley; [3]Plant and Microbial Biology Department, University of California, Berkeley; [4]Department of Bioengineering, University of California, Berkeley; [5]Lawrence Berkeley National Laboratory; and [6]Computer Science Department, University of Chicago, Chicago, Ill.

Project Goals: We are awash in proteins discovered through high-throughput sequencing projects. As only a minuscule fraction of these have been experimentally characterized, computational methods are widely used for automated annotation. Unfortunately, these predictions have littered the databases with erroneous information, for a variety of reasons including the propagation of errors and the systematic flaws in BLAST and related methods. In collaboration with Michael Jordan's group, we have developed a statistical approach to predicting protein function that uses a protein family's phylogenetic tree, as the natural structure for representing protein relationships. We overlay on this all known protein functions in the family. We use a model of function evolution to then infer the functions of all other protein functions. Even our initial implementations of this method, called SIFTER (statistical inference of function through evolutionary relationships) have performed better than other methods in widespread use. We are presently making numerous improvements to the underlying SIFTER algorithm and enhancing its ability to work on a wide range of data.

It is now easier to discover thousands of protein sequences in a new microbial genome than it is to biochemically characterize the specific activity of a single protein of unknown function. Through metagenomic analysis, next-generation sequencing heralds unprecedented opportunities for understanding the environmental microbiota. A single experiment alone, the Global Ocean Sampling study, more than doubled the number of known protein sequence entries. However, despite this large body of new sequence information, functional annotation remains a major challenge. Molecular functions of proteins in the novel genomes continue to be discovered, in large part by homology to those experimentally characterized in model organisms.

Typically, protein function annotation involves finding homologs of a protein sequence, followed by database queries and computational techniques to predict function from the annotated homologs. These methods rely on the principle that proteins from a common ancestor may share a similar function. However, most protein families have sets of proteins with different functions and therefore traditional bioinformatics approaches are unable to reliably assign the appropriate function to unannotated proteins. Currently, protein function databases have a large proportion of erroneously annotated proteins, where the incorrect annotations were either derived using an imprecise computational technique or inferred using another incorrect annotation.[1-4]

We have proposed integrating available functional data using the evolutionary relationships of a protein family, and we implemented this method in the program SIFTER (Statistical Inference of Function Through Evolutionary Relationships). The SIFTER methodology uses a statistical graphical model to compute the probabilities of molecular functions for unannotated proteins. Currently, SIFTER takes as input a reconciled phylogeny and a set of annotations for some of the proteins in the protein family. We incorporate known information about function by computing the probability of each of the candidate functions for the proteins in the tree with available functional evidence from the GOA database. The candidate molecular functions are represented as a boolean vector, where initially the probability associated with each candidate function is a function of the set of annotations for that protein and their corresponding evidence types (e.g., experimental, electronic). From this reconciled phylogeny with sparse observations, SIFTER computes the posterior probability of each molecular function for all proteins in the family using a simple statistical model of protein function evolution.

We tested the performance of SIFTER on three different protein families: AMP/adenosine deaminases, sulfotransferases and Nudix hydrolases with cross-validation experiments. SIFTER's performance was compared with three other function prediction algorithms: BLAST, GOtcha and Orthostrapper, and SIFTER was shown to outperform the other methods. In addition, on a genome-wide scale we used SIFTER to annotate the experimentally characterized proteins from *Schizosaccharomyces pombe*, based on the annotations from 26 other fungal genomes. The newest version of SIFTER implements a faster method for calculating the posterior probabilities, and this improvement, together with a more general evolutionary model make SIFTER applicable on large and functionally diverse protein families and on genome-scale function annotation.

The development of SIFTER is an ongoing project and a new version of the program is now available (manuscript under review). We are currently testing SIFTER for metagenomic sequences with the acid mine drainage datasets from Jill Banfield. In the near future, we are planning to expand our analysis to other metagenomic datasets, such as the termite gut datasets from the JGI. We also use SIFTER to annotate enzymes from chlorite dismutase and perchlorate reductase families, in order to identify species that are capable of perchlorate reduction. Furthermore, we are validating SIFTER predictions experimentally using the large and extremely diverse Nudix family of hydrolases as a test bed.

References

1. Brenner SE 1999 *Trends Genet.* 15 132-3
2. Galperin MY and Koonin EV 1998 *In Silico Biol.* 1 55-67
3. Jones CE, Brown AL and Baumann U 2007 *BMC Bioinformatics.* 8 170
4. Schnoes AM, Brown SD, Dodevski I and Babbit PC 2009 *PLoS Comput Biol.* 5 e1000605

This project has been funded with DOE grant BER KP 110201.

Structural Biology, Molecular Interactions, and Protein Complexes

240

Neutron Protein Crystallography Station User Facility

S. Zoe Fisher, Andrey Kovalevsky, Marat Mustyakimov, Marc-Michael Blum, Benno P. Schoenborn, Mary Jo Waltman, and **Paul Langan*** (Langan_paul@lanl.gov)

Bioscience Division, Los Alamos National Laboratory, Los Alamos, N.M.

Project Goals: PCS is a high-performance neutron beamline that forms the core of a BER-funded experimental User capability at Los Alamos Neutron Science Center (LANSCE) for investigating the structure and dynamics of proteins, biological polymers, and membranes.

Neutron diffraction is a powerful technique for locating hydrogen atoms, which can be hard to detect using X rays, and therefore can provide unique information about how biological macromolecules function and interact with each other and smaller molecules. This unique User capability is being used to investigate several enzymes that are important to USDA and DOE Genome Science program missions in renewable energy and the environment, with a view to understanding their detailed catalytic mechanisms. This new information is then being exploited to manipulate their performance and use. Neutron diffraction has also been crucial in revealing the structures and hydrogen bond arrangements in naturally occurring cellulose in lignocellulosic biomass and how they are rearranged by pretreatments to enhance conversion to biofuels. This information has led to the optimization of pretreatments to improve their cost-efficiency.

PCS Users have access to neutron beam time, deuteration facilities, protein expression and substrate synthesis with stable isotopes, a purification and crystallization laboratory, and software and support for data reduction and structure analysis. A HomeFlux X-ray system has been recently purchased that will allow users to collect X-ray data from the same samples used for neutron diffraction. The PCS beamline exploits the pulsed nature of spallation neutrons and a large electronic detector to efficiently collect wavelength-resolved Laue patterns using time-of-flight techniques. We encourage potential users to communicate with us before applying for beam time for technical guidance and help with proposal preparation.

For technical information about the PCS and experimental requirements contact Zoe Fisher (505) 665-4105 zfisher@lanl.gov or Paul Langan (505) 665 8125 langan_paul@lanl.gov

Proposal Submission: Proposals must be submitted using the process on the LANSCE website. To access the proposal submission site, go to the LANSCE home page, http://lansce.lanl.gov/. On this page click the tab “Lujan Center’ and then the link ‘Submit a Proposal.This will take you to the on-line submission system. Detailed instructions for preparing the proposals can be found on the proposal submission sites under “Step-by-Step Guide to Submitting an Online Proposal.”

241

The Berkeley Synchrotron Infrared Structural Biology (BSISB) Program Overview

Hoi-Ying N. Holman[1,2]* (hyholman@lbl.gov), Hans A. Bechtel,[1,3] Rafael Gomez-Sjoberg,[1,4] Zhao Hao,[1,2] Ping Hu,[1,2] Michael C. Martin,[1,3] and Peter Nico[1,2]

[1]Berkeley Synchrotron Infrared Structural Biology Program, [2]Earth Sciences Division, [3]Advanced Light Sources, and [4]Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: The Berkeley Synchrotron Infrared Structural Biology (BSISB) program is a national user facility for infrared spectromicroscopy and chemical microcharacterization of living cells.

The Berkeley Synchrotron Infrared Structural Biology (BSISB) program is a national user facility for infrared spectromicroscopy and chemical microcharacterization of biological systems. BSISB was initiated in 2010 to maintain a forefront research facility for infrared and optical characterization of chemistry in living cells with state-of-the-art instrumentation and expertise. The BSISB program has developed an integrated microfluidic synchrotron infrared (SIR) spectromicroscopy platform, which is a technique that is ideal for tracking the chemical composition and reactions in living cells during their adaptive responses to internal or external stimuli and perturbations. The BSISB program is also developing visible (VIS) hyperspectral/fluorescence microscopy approaches for simultaneously tracking changes in cellular morphology, structure, and other biological processes such as gene expression and signaling during SIR experiments. This new BSISB development of live-cell chemical biological imaging technologies will also be aided by a new generation of microfluidics platform. Our technological research and development effort will be accelerated by BSISB participating scientists with wide ranging research projects of bioenergy, medical, and environmental studies