U.S. DEPARTMENT OF
ENERGY

DOE/SC-0110

Joint Meeting

Genomics:GTL Awardee Workshop VII

and

USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Workshop 2009

Bethesda, Maryland
February 8–11, 2009

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

http://genomicsgtl.energy.gov

Prepared by
Genome Management 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

Contents

Great Lakes Bioenergy Research Center (GLBRC)

Bioenergy Sciences Center (BESC)

Biofuels > Analytical and Imaging Technologies for Studying Lignocellulosic Material Degradation

Systems Biology and Metabolic Engineering Approaches for Biological Hydrogen Production

SBIR Bioenergy Research

Systems Environmental Microbiology

The Virtual Institute of Microbial Stress and Survival VIMSS:ESPP

The *Shewanella* Federation

Caulobacter Systems Understanding

AMD Community Systems Understanding

Molecular Interactions and Protein Complexes

The MAGGIE Project

The Center for Molecular and Cellular Systems

Validation of Genome Sequence Annotation

Computing Resources and Databases

Communication.....187

Addendum of late abstracts available page 215. Late abstracts not indexed.

Introduction to Workshop Abstracts

Genomics:GTL Goal and Objectives

Ultimate Scientific Goal
Achieve a predictive, systems-level understanding of plants, microbes, and biological communities, via integration of fundamental science and technology development, to enable biological solutions to DOE mission challenges in energy, environment, and climate.

Objective 1: Determine the genomic properties, molecular and regulatory mechanisms, and resulting functional potential of microbes, plants, and biological communities central to DOE missions.

Objective 2: Develop the experimental capabilities and enabling technologies needed to achieve a genome-based, dynamic systems-level understanding of organism and community functions.

Objective 3: Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance the understanding, prediction, and manipulation of complex biological systems.

Abstract Organization

Genomics:GTL program abstracts and posters are organized according to the following research areas important to achieving the ultimate GTL scientific goal and objectives.

Systems Biology for DOE Energy and Environmental Missions

Bioenergy

- Biofuels: Bioenergy Research Centers
- Biofuels: USDA–DOE Plant Feedstock Genomics for Bioenergy
- Biofuels: Analytical and Imaging Technologies for Studying Lignocellulosic Material Degradation
- Systems Biology and Metabolic Engineering Approaches for Biological Hydrogen Production
- SBIR Bioenergy Research

Systems Environmental Microbiology

Systems Biology Strategies and Technologies for Understanding Microbes and Microbial Communities

Genomic and Proteomic Strategies

Molecular Interactions and Protein Complexes

Validation of Genome Sequence Annotation

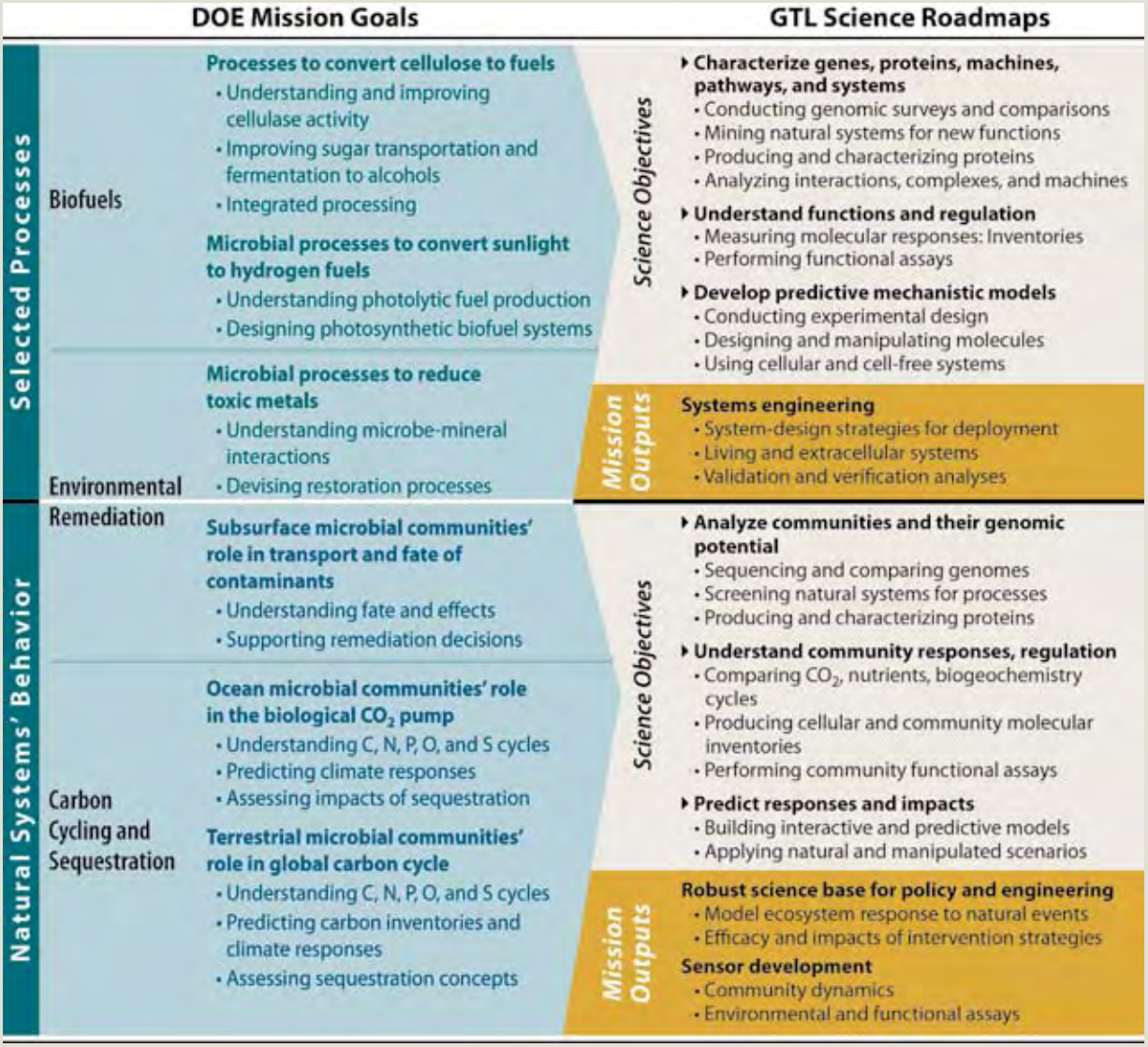
Computing Resources and Databases

Communication

Ethical, Legal, and Societal Issues

The following table is a summation of how GTL science and DOE missions align (*DOE Genomics:GTL Roadmap: Systems Biology for Energy and Environment*; October 2005, p. 40). (GenomicsGTL.energy.gov)

Summary Table. GTL Science Roadmap for DOE Missions



A capsule summary of systems being studied, mission goals that drive the analysis, generalized science roadmaps, and outputs to DOE missions. To elucidate design principles, each of these goals entails the examination of thousands of natural primary and ancillary pathways, variants, and functions, as well as large numbers of experimental mutations.

Systems Biology for DOE Energy and Environmental Missions

Bioenergy

Biofuels > Bioenergy Research Centers

Joint Bioenergy Institute (JBEI)

GTL

Systematic Characterization of Glycosyltransferases Involved in Plant Cell Wall Biosynthesis

Henrik Vibe Scheller[1]* (hscheller@lbl.gov), Ai Oikawaa, Lan Yin,[1,2] Eva Knoch,[1,2] Naomi Geshi,[2] Carsten Rautengarten,[1] Yuzuki Manabe,[1] and Chithra Manisseri[1]

[1]Feedstocks Division, Joint Bioenergy Institute, Emeryville, Calif.; and [2]Dept. of Plant Biology and Biotechnology, University of Copenhagen, Denmark

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

The goals of the project are to provide a detailed understanding of the enzymes that are responsible for biogenesis of the plant cell wall and develop a knowledgebase to enable generation of crop plants with improved properties as feedstocks for biofuel production.

Plant cell walls are composed mainly of polysaccharides and production of biofuels from biomass requires decomposition of the polymers. Many of the polymers are recalcitrant to degradation and they are composed of sugars that are not optimal for fermentation. Better understanding of the biosynthesis of the cell wall polysaccharides may enable development of crops with improved properties as biofuels feedstocks. Despite rather detailed information on the structure of the cell wall polysaccharides, little is known about their biosynthesis. The key enzymes are glycosyltransferases (GTs) and plants need a large number of GTs to synthesize the complex polysaccharides present in the walls. However, only a few GTs have had their activity demonstrated. In *Arabidopsis thaliana*, approximately 450 GT genes have been identified based on their sequence and deposited to the CAZy database (www.cazy.org). We have cloned a large number of these GTs in Gateway vectors in order to heterologously express the GTs and characterize their activity. Systematic analysis of the GTs is in progress and results will be presented. Agrobacterium-mediated transient expression in *Nicotiana benthamiana* has a high success rate for expression of GTs.

An alternative way to elucidate the function of GTs and other biosynthetic enzymes is to study the effect of down-regulating or inactivating the corresponding genes. This approach is often hampered by the overlapping function of many GTs. Generation of mutants that are affected in several homologous genes can overcome this limitation. In other cases, GTs are functioning in complexes that contain several different polypeptide subunits. The CslD family of proteins provides examples of both redundancy in function and of protein complexes.

GTL

Analysis of Putative Feruloyltransferase Transcript Levels and Cell Wall Composition During Rice Development

Laura E. Bartley[1,2]* (LEBartley@lbl.gov), Chithra Manisseri,[2] Brian Williams,[1,2] Dawn Chiniquy,[1,2] Jesper Harholt,[2] Henrik Vibe Scheller,[2] and Pamela C. Ronald[1,2]

[1]Dept. of Plant Pathology, University of California, Davis, Calif.; and [2]Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Identification of genes encoding cell wall modifying enzymes has applications in human and animal nutrition, plant defense against pathogens, and biofuel production efficiency. For plants with type II cell walls, such as rice and other grasses, glucuranoarabinoxylans (GAX) are a major component of the hemicellulose found in both primary and secondary walls. Ferulic acid, a phenolic compound, is added to the O-5 of arabinosyl units of GAX. Ferulic acid residues can covalently crosslink arabinosyl residues through dimer formation, or serve as attachment points between

GAX chains and lignin, providing rigidity to the cell wall matrix. Therefore, reducing the ferulic acid content of Type II cell walls may improve feedstock deconstruction for production of lignocellulosic biofuels. The genes that encode feruloyltransferases (FTs) have not yet been identified. However, Mitchell and colleagues (Plant Physiology 2007) have recently predicted that FTs may be encoded by a clade of genes from the CoA acyl transferase superfamily based on their greater EST abundance in grasses compared with dicots. Ratios of cell wall components, including GAX and ferulic acid, are dynamic, adjusting to allow for elongation while plants are young, and rigidity, as they mature. To correlate cell wall composition and gene expression during rice development, we have collected RNA and cell wall samples from various rice tissues throughout development. We are conducting quantitative RT-PCR to amplify transcripts from the putative FT genes identified by Mitchell et al. Cell wall sugar and phenolic composition will be examined by HPLC analysis. This study will determine how rice cell wall composition shifts to accommodate the changing needs of the plant, providing valuable data for further rice genomics analyses and evidence to test the hypothesis of Mitchell et al. regarding the putative FTs. Genes from the 20-member gene family that show correlation with ferulic acid accumulation will be targeted for functional studies toward improving the deconstruct-ability of grass cell walls for biofuels.

GTL

The Utilization of *Arabidopsis* Genetic Variants to Understand Cell Wall Structure and Biosynthesis

Joshua L. Heazlewood* (JLHeazlewood@lbl.gov), Katy M. Christiansen, Dominique Loque, and A. Michelle Smith

Feedstocks Division, Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

The process of plant cell wall biosynthesis involves a complex series of biochemical processes involving many hundreds of proteins. Determining the function of a specific gene through functional genomics techniques have proved problematic due to genetic redundancies and undetectable changes. Common techniques have used mutant collections in forward genetic screens or reverse genetics to directly target and disrupt genes of interest. Since such techniques are heavily reliant on some phenotypic discrimination to asses gene function, the absence of a measureable difference often results in little useful information. Furthermore the complete absence of many genes produced by such techniques results in a lethal phenotype as the gene is absolutely required for normal function of the plant. A more subtle approach utilizes genetic differences in naturally occurring variants to provide important information about gene function and genetic diversity. In collaboration with the Joint Genome Institute we have sequenced two *Arabidopsis* accessions (Bay-0 and Sha-0) that have previously been shown to have measureable differences in Ara-Rha ratios in cell wall extracts. This genetic information and the utilization of recombinant inbred lines (RIL's) will be used to map QTL's identified in these and other *Arabidopsis* accessions to identify loci that contribute to functional differences in plant cell walls.

GTL

Selection, Cloning and Functional Characterization of Rice-Diverged, Cell Wall-Related Glycosyltransferases

Miguel Vega-Sánchez* (mevegasanchez@ucdavis.edu), Laura Bartley, Peijian Cao, and Pamela Ronald (pcronald@ucdavis.edu)

Dept. of Plant Pathology, University of California, Davis, Calif.; and Joint BioEnergy Institute, Feedstocks Division, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Understanding plant cell wall biosynthesis is crucial for the development of the next generation of biofuels derived from lignocellulosic material. Current limitations in the harvest of fermentable sugars from cellulose derive from the inherent recalcitrance of plant cell walls. Basic knowledge of how the structure and composition of the cell wall can be modified to obtain biomass suitable for efficient and economically viable biofuel production is needed. Glycosyltransferases (GTs) are a large, multifamily class of enzymes that form glycosidic bonds between donor nucleotide sugars and acceptor substrates. Among the GTs are the enzymes responsible for the synthesis of important cell wall polysaccharides, including cellulose, hemicellulose and callose. However, the function, substrate specificity and biochemical activity of the majority of GTs are unknown. Many aspects of grass and other commelinoid monocot cell walls are distinct from that of better-studied dicots. As many of the preferred feedstocks for future bioenergy initiatives include grasses such as switchgrass and *Miscanthus*, the model species of choice to advance our understanding of monocot-specific aspects of the cell wall is rice (*Oryza sativa*). Rice was the first grass species to have its full genome sequenced and abundant

functional genomic resources have accumulated in recent years. To identify potential targets involved in cell wall synthesis, we mined the recently created and publicly available rice GT phylogenomic database (http://ricephylogenomics.ucdavis.edu/cellwalls/gt/) to select a list of 33 rice-diverged GTs with high expression in above-ground, vegetative tissues. Cloning of these genes is underway, and will be followed by the creation of a variety of expression constructs for functional analysis in both transgenic plants and heterologous systems for protein expression, biochemical activity determination and protein-protein interaction network generation. We have ordered insertion and activation-tagged lines from worldwide repositories for cell wall composition/modification analyses as well as phenotypic characterization. Among this group of genes, we will study in detail a subset within GT family 2, the grass-specific Cellulose synthase like (Csl)-F and H sub-families, by generation of gene family knockdowns using artificial miRNAs. We will characterize other putative GTs from unknown families as well. We anticipate that by focusing on grass-specific GTs involved in cell wall synthesis and modification in plant tissues important for the production of lignocellulosic-derived biofuels, we can facilitate the improvement of genetic traits in second-generation bioenergy crops such as switchgrass.

GTL

Monolignol Transporters and Cell Wall Oxidases Screens

Fan Yang* (Fyang@lbl.gov), Nakako Shibagaki, and Dominique Loque

Joint BioEnergy Institute, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Currently, biofuels, such as ethanol are produced largely from starch that comes from grains, but it represents only a little proportion of sugar polymer availability on Earth. Large quantities of sugar from polysaccharides that are not utilized thus far are cellulose and hemicellulose, which are the main constituent of plant cell walls. The energy efficiency of starch-based biofuels is however not optimal, while plant cell walls (lignocellulose) represent a huge resource for BioEnergy since plant cell walls are composed of 95-70% of sugar. The rest of the plant cell walls (5-30%) is mainly composed of lignin, which is a very strong phenolic polymer recalcitrant to degradation and inhibits efficient extraction of sugars from the cell wall thus prevents cost-effective lignocellulosic ethanol production. Unfortunately, lignin provides compressive resistance to plant cells and cannot simply be genetically removed without incurring deleterious consequences on plant productivity. Lignin gives a strong structural support to the plant but also protects the plant against biotic (e.g. pathogens) and abiotic stress (e.g. UV, wind). Therefore, it is important to understand to which extent plant lignin content and composition can be modified in useful ways without deleterious consequences to plant growth and development. Since lignin cannot be removed, a better control of lignin deposition and cross-linking within the cell wall may increase sugar recovery from the cell wall polysaccharides. Another approach would be to replace the "hard bounds" (e.g. ether, carbon bonds) in the lignin polymer with easily cleavable ones (e.g., amide or ester bonds) and thus would facilitate removal of lignin with processes that are much more cost effective than current processes.

In order to be able to modify lignin deposition, polymerization and cell wall cross-linking, a better knowledge of the enzymes (oxidases) that participates in the polymerization of lignin and cross-linking to other cell wall components is required. Similarly, to be able to control monolignol export into the appoplast, monolignol transporters need to be identified and characterized. Therefore, we are currently developing a strategy using yeast complementation to try to indentify protein involved in lignin deposition. Plant cDNA libraries, a complete *Arabidopsis* MDR transporters library and selected genes will be heterologous expressed in yeast and tested for their ability to detoxify phenolic compounds by export or polymerization.

GTL

Starting Point for Enzymatic Hydrolysis for Cellulose: Enzyme Engineering of Glycoside Hydrolase-5 Endoglucanases

Blake A. Simmons[1,2]* (basimmons@lbl.gov), Zhiwei Chen,[1,2] Supratim Datta,[1,2] Danielle T. Ercek,[1] Hanbin Liu,[1,2] Josh I. Park,[1,2] Huu Tran,[1,2] Swapnil R. Chhabra,[1] Masood Z. Hadi,[1,2] and Rajat Sapra[1,2]

[1]Joint BioEnergy Institute, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [2]Sandia National Laboratories, Livermore, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Enzymes are catalysts that have a set of specific and well defined characteristics like substrate, pH, temperature and kinetics etc. The vast amount of available genomic data shows that orthologous genes code for proteins that vary in these characteristics and the changes in properties at the protein level can be traced to sometimes minor changes like substitutions in the sequence at the molecular level. This observation that changes in the sequence leads to changes in

characteristics can be used to design new and novel protein variants using two major approaches—directed evolution and rational design. We propose to use these techniques to engineer enzymes with improved activity and characteristics better suited to optimal conditions for cellulosic biomass deconstruction. Directed evolution of enzymes is a powerful technique that takes advantage of the Darwinian process of natural selection on a high throughput lab scale for the generation of protein mutants, called variants, which are then screened and selected for improved desirable traits as compared to the parent protein characteristics. At the core of the technique is the principle that incremental changes acquired either through mutagenesis or via recombination lead to a better variant when selection pressure is applied; the selection pressure can be any of the characteristics that are sought to be improved upon like activity, kinetics, pH or temperature stability and, in some cases, different substrates and novel reactions. The advantage for using the directed evolution approach is that there is no requirement for the availability of an extensive data set of orthologs as a starting point; only a gene sequence and a screening method for 'evolving' a protein function is needed. Therefore, this technique can be employed in all the enzymes currently known to be involved in cellulosic biomass deconstruction- cellulases for which there are a large number of gene sequences available, xylanases and ligninases etc. for which there is scant gene and enzyme data available in the database.

GTL

A Comparative Study of Dilute Acid and Ionic Liquid Pretreatment of Biomass and Model Lignocellulosics

Chenlin Li[1,2*] (Cli@lbl.gov), Rohit Arora,[1,2] Chithra Manisseri,[1] Ken Vogel,[3] Blake Simmons,[1,2] and Seema Singh[1,2]

[1]Joint BioEnergy Institute, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Emeryville, Calif.; [2]Energy Systems Dept., Sandia National Laboratories, Livermore, Calif.; and [3]United States Department of Agriculture, Lincoln, Nebr.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Lignocellulosic biomass has great potential to serve as a low-cost and abundant feedstock for bioconversion into fermentable sugars, which can be subsequently fermented into biofuel. There are several obstacles that remain to be solved in this conversion process, specifically lignin content, crystalline cellulose structure and the presence of ester linkages between lignin and hemicellulose in the plant cell wall limit the enzymatic accessibility of the recovered polysaccharides for efficient saccharification into sugars. Various physical and chemical pretreatment methods are currently employed to break down the biomass recalcitrant structures, and increase their susceptibility to enzymes. Among these techniques, dilute acid pretreatment has been shown as a leading pretreatment process. However, dilute acid hydrolysis can lead to degradation products that are often inhibitory and significantly lower the overall sugar yields. Glucose and xylose degradation products that result from the pretreatment methods include hydroxymethylfurfural (HMF) and furfural, which produce levulinic and formic acids, respectively, which inhibit the subsequent fermentation of sugars to ethanol. Recently, ionic liquids have demonstrated great promise as efficient solvents for biomass with easy recovery of cellulose upon anti-solvent addition. However, to date, no comprehensive side-by-side comparative analysis has been conducted in order to evaluate the dilute acid and ionic liquid biomass pretreatment processes. In this study, we compare ionic liquid and dilute acid pretreatments acting on switchgrass with numerous analytical techniques to gain a better understanding of both techniques and the resultant saccharification yields of fermentable sugars.

GTL

Understanding Ionic Liquid Pretreatment of Lignocellulosic Biomasses

Seema Singh[1*] (SSingh@lbl.gov), Rohit Arora,[1] Chithra Manisseri,[1] Chenlin Li,[1] Henrik Vibe Scheller,[1] Ken Vogel,[2] and Blake Simmons[1]

[1]Joint BioEnergy Institute, Emeryville, Calif.; and [2]United States Department of Agriculture, Lincoln, Nebr.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Pretreatment of biomass is essential for breaking apart highly ordered and crystalline plant cell walls and loosening the lignin and hemicellulose conjugation to cellulose microfibrils, thereby facilitating enzyme accessibility and adsorption and reducing costs of downstream saccharification processes. Recent reports[1,2] have shown very high yields at very low enzyme loadings. However, pretreatment still remains one of the most costly steps in lignocellulosic biofuel production. Ionic liquids are novel solvents showing great promise for lignin and cellulose solubilization. Instant rejection of dissolved polysaccharides upon addition of anti-solvent shows promise for recyclability in addition to other desired attributes like low volatility, non-flammability and thermal stability. Although ionic liquids have been shown to be very effective in cellulose solubilization[3,4], the disposition of hemicellulose and lignin are not fully understood. The

* Presenting author

aim of our research is to develop a fundamental understanding of ionic liquid pretreatment by monitoring and analyzing process streams. To that end, we have employed HPAEC, XRD, FTIR, NIR, and SEM to study the impact of ionic liquid pretreatment on switchgrass and corn stover. We will present the results from these measurements in the context of developing and selecting optimized ionic liquid pretreatment conditions for selective depolymerization of either cellulose or lignin, whereby fractionation of different cellulosic and lignin components could be realized.

References

1. Teymouri, F., Laureano-Perez, L., Alizadeh, H., & Dale, B. E. (2005). *Bioresource Technology,* 96, 2014–2018.
2. Kim, T. H., Kim, J. S., Sunwoo, C., & Lee, Y. Y. (2002). In *24th Symposium on Biotechnology for fuels and Chemicals*.
3. Dadi et al, *Applied Biochemistry and Biotechnology*, 136-140, 407.
4. Singh, S. & Simmons, B, In-Situ Examination of Biomass Dissolution and Cellulose Regeneration Enabling Cellular level Insight of Ionic Liquid Pretreatment Process (manuscript in preparation).

GTL

JBEI Microbial Communities Deconstruction Research Activities

Terry C. Hazen* (TCHazen@lbl.gov), Phil Hugenholtz, Steve Singer, Jean VanderGheynst, Patrik M. D'haeseleer, Michael P. Thelen, Kristin DeAngelis, Amitha Reddy, Martin Allgaier, Julian Fortney, Gary Andersen, Todd DeSantis, Eoin Brodie, Cindy Wu, Dara Goodheart, Mary Firestone, Whendee Silver, and Blake Simmons

Joint BioEnergy Institute, Deconstruction Division, Microbial Communities Dept., Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

There exist several unexplored microbial communities that are promising sources of biocatalysts and metabolic pathways for hydrolysis of polysaccharides in plant cell walls (lignocellulose), such as those found on the tropical rain forest floor and within compost. The compost environment is nutrient rich, whereas the rain forest floor is nutrient deficient. Our approach is to examine both communities to increase the likelihood of discovering a variety of novel pathways for the biological deconstruction of lignocellulose. The tropical rain forest has developed metabolic pathways that quickly convert the available biomass into useful metabolites as evidenced by the scarcity of litter on the forest floor despite high net primary productivity. These communities therefore have great potential to provide enzymes with very high specific activities and rapid kinetics of biomass degradation. In contrast, composting is an industrial process managed under thermophilic conditions. Microbial communities in compost are dynamic and vary based on the extent of decomposition of the compost, moisture, temperature, and oxygen concentration. Studying compost communities, in particular composts enriched in lignocellulosic biomass (greenwaste composts), will facilitate the discovery of novel enzymes and pathways amenable to certain high-temperature, high ionic strength, and non-neutral pH biomass pretreatment conditions, as well as insights into how such pathways are regulated for optimal lignocelluose degradation in their respective environments. Forests, and in particular tropical rain forests, have evolved to obtain the maximum nutrients from leaf litter. Thus, over many millions of years, the soil microbial communities in these environments have evolved to recycle all biomass that results from the episodic and changing forest environment. For example, the microbial communities present in the wet tropical forest soils of Puerto Rico have optimized the conversion of lignin and cellulose to methane and methanol through 40-60 million years of evolutionary selection for rapid and complete decomposition of plant material. Wet tropical forests typical of those in Puerto Rico have some of the highest known rates of plant productivity on Earth and concomitantly high microbial rates of biomass breakdown. Due to abundant rainfall and high carbon availability, these soils are frequently anaerobic leading to methane production during decomposition. Studies have confirmed that because of this frequent anaerobiosis, prokaryotes assume a primary role in macromolecular decomposition. Wet tropical soils will provide fertile ground for the discovery of novel lignin and cellulose degrading enzymes, and for new and unique consortia capable of effectively functioning under a wide range of environmental conditions. The aerobic-anaerobic cycling that naturally occurs over time and space in these wet forest soils has made selected bacterial and archaeal communities extremely effective in converting lignocellulose to methane. In composting prokaryotes are dominant in the early stages of decomposition, while both prokaryotes and eukaryotes play a role later in the process. Selecting incubation conditions that enrich both groups is an important part of our experimental approach. The Microbial Communities Department will identify and isolate key lignocellulolytic enzymes and metabolic pathways that will be delivered to the Deconstruction Division enzyme engineering group and the Fuels Synthesis Division, respectively of JBEI. During the past year this group has constructed and tested a new MycoChip microarray for rapid identification of more than 11,620 fungal taxa. Early in the year we placed bags of switch grass and lignin baited "bug traps" in the soil of the tropical rain forest to enrich for lignocellulose degraders. In only 1 week we observed increases in phenol oxidase and peroxidase. We have also observed a very high microbial diversity in these samples. Respirometer studies of switch grass with tropical rain forest soil revealed very high rates of production of CO_2, CH_4 and H_2S from the inoculated switchgrass and more then 156 different taxa from the soil alone. We also established optimal wetting conditions for

decomposition of corn stover and switchgrass inoculated with green waste compost and incubated under aerobic and thermophilic conditions. We determined that there were more than 2,602 bacterial taxa associated with switchgrass inoculated with compost. Preliminary results on Titanium 454 pyrosequencing of one tropical rain forest soil sample revealed more than 2,700 cellulases, hemicellulases, and ligninases. Similar data on a compost sample are currently being analyzed. We intend to characterize the enzyme repertoire present in the metagenomic samples, and correlate that with differences in activity and community composition between compost and tropical rain forest soil.

GTL

Discovery and Optimization of Lignocellulolytic Bacteria From Puerto Rican Rainforest Soils

Kristen M. DeAngelis* (KDeAngelis@lbl.gov), Whendee L. Silver, Julian Fortney, and Terry C. Hazen

Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Tropical soils in Puerto Rican rain forests have some of the highest decomposition rates recorded in the world, with almost total mass lost in decomposing plant material within one year. These soils are capable of deconstructing biofuel plant materials to basic components, like ethanol, methane or methanol. Rapidly fluctuating redox conditions are characteristic of the highly weathered soils of upland humid tropical forests, which are dominated by Fe-oxide mineralogy and have relatively low sulfate availability. The frequent episodes of anoxic conditions make it likely that these decomposing consortia are primarily bacteria, not fungi as are usually observed in temperate systems. Previous lab incubation under fluctuating redox conditions permitted simultaneous methanogenesis, N_2O production, and iron reduction, all accompanied by steady CO_2 production. The objective of this research is to define field conditions that result in characteristic high methane production in Puerto Rico forest soils from decomposing plant materials, and determine whether different microbial communities break down different plant materials. Towards this end, we designed a field experiment and accompanying laboratory incubations that would allow us to investigate the rates, controls and mechanisms of switchgrass decomposition in tropical rainforest soils. In June of 2008, we buried litterbags filled with switchgrass in four different forest types in the Luquillo LTER, located at the El Yunque National Forest in Puerto Rico, USA. The four forest types vary from more aerobic soils, warmer temperatures and annual precipitation on the order of 1,000 mm, to fluctuating redox soils, to mostly anaerobic soils, cooler soil temperatures and annual precipitation that can exceed 4,000 mm.

The experimental design included 4 field sites, 6 time points, and bags buried in pairs, one for. At each of 6 time point, litter bags and soil are collected from the field and assayed for microbial community analysis using 16S ribosomal DNA PhyloChip, potential enzyme activity (β-glucosidase, endoglucanase xylosidase, chitinase, phenol oxidase and peroxidase), and mass loss as indicators of decomposition. In the driest site, which we expect to also have the highest rates of decomposition, we also buried biosep beads baited with lignin (using unbaited beads as controls) as bug traps to identify and isolate microbes specifically able to decompose lignin. This site was also instrumented with oxygen sensors to measure oxygen levels in soil on an hourly basis over the course of this year-long incubation, and ultimately to correlate decomposition, enzyme activity and microbial community composition with oxygen availability at the end of the experiment. Concomitantly with the field experiment, we are using fresh soil to inoculate mini-reactors with dried ground switchgrass and incubate anaerobically to enrich for lignocellulose-degrading organisms. The initial inoculation of rain forest soil with switchgrass resulted in significant CO_2, CH_4 and H_2S production compared to uninoculated, anaerobic soil incubations, as well as a substantial change in microbial community composition. Switchgrass amendment resulted in significant change in 147 taxa compared to the 1847 detected in the soils. With switchgrass addition to soil, Archaea, methanogens, enteric bacteria, *Bacilli* and *Clostridia* were significantly increased, while *Acidobacteria*, *Burkholderia* and *Verrucomicrobial* were significantly reduced in the microbial community. Further passages of the soil microbial community with switchgrass as the sole carbon source has resulted in a low-richness, anaerobic microbial community capable of efficiently converting switchgrass to methane and carbon dioxide as well as depolymerizing cellulose, hemicellulose, and lignin in the process.

GTL

Metagenomic Characterization of Compost and Rain Forest Soil Microbial Communities

Martin Allgaier[1]* (MAllgaier@lbl.gov), Amitha Reddy,[1,2] Jean VanderGheynst,[2] Alex Copeland,[3] Victor Kunin,[3] Patrik D'haeseleer,[4] Kristen DeAngelis,[1,5] Julian Fortney,[1,5] Dylan Chivian,[1,5] Paramvir S. Dehal,[5] Blake Simmons,[1,6] Terry C. Hazen,[1,5] and Phil Hugenholtz[1,3] (PHugenholtz@lbl.gov)

[1]Joint BioEnergy Institute, Emeryville, Calif.; [2]Dept. of Biological and Agricultural Engineering, University of California, Davis, Calif.; [3]DOE Joint Genome Institute, Walnut Creek, Calif.; [4]Lawrence Livermore National Laboratory, Livermore, Calif.; [5]Lawrence Berkeley

National Laboratory, Berkeley, Calif.; and [6]Sandia National Laboratories, Livermore, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Microorganisms are a promising source of novel carbohydrate-active enzymes (cellulases, hemicellulases, and lignases) since they are primarily responsible for plant biomass degradation in nature. However, most carbohydrate-active enzymes used by industry come from only a few model organisms. To identify new lignocellulolytic enzymes we shotgun sequenced genomic DNA from two sources: a sample of pristine Puerto Rican rain forest soil and a sample obtained from a solid state fermentation experiment in which switchgrass was incubated 30 days in a lab under mesophilic and thermophilic conditions after inoculating with green waste compost from an industrial facility. Both ecosystems display high rates of plant biomass degradation and are therefore prime targets for novel carbohydrate-active enzyme discovery.

454-Titanium pyrosequencing was used to generate metagenome data sets from the two samples resulting in a total of 1,412,492 reads (rain forest: 863,759; compost: 548,733) with an average read length of 424 bases. Reads were quality filtered and trimmed in preparation for comparative analyses with the metagenome analysis tools IMG/M and MicrobesOnline. The complexity of the rain forest soil metagenome precluded assembly, so sequence data were analyzed unassembled. However, for the compost sample, significant assembly occurred resulting in contigs up to 50 kb in length.

Preliminary comparative analysis of a fraction of the rain forest soil sequence data revealed more than 2,700 cellulases, hemicellulases, and ligninases including glycoside hydrolases as well as glycosyl transferases, representing ~1% of all predicted protein-coding genes (e.g. compared to 1.2% or 0.03% identified in metagenomic data sets from termite guts or silage surface soil, respectively). The enzyme repertoires present in the two metagenome data sets will be further analyzed to identify new deconstruction enzymes and compared to assess differences in activity and community composition between compost and rain forest soil.

GTL

The Fuels Synthesis Division of the Joint BioEnergy Institute (JBEI)

E. Baidoo, H.R. Beller, R. Chan, S. Chhabra* (SRChhabra@lbl.gov), H. Chou, R. Dahl, Z. Dmytriv, M. Dunlop, C. Fortman, D. Garcia, H. Garcia Martin, J. Gilmore, J. Gin, E.-B. Goh, J. Haliburton, T. Ham, C. Joshua, Y. Kang, R. Krupa, S.K. Lee, T.S. Lee, C. Liu, A. McKee, A. Mukhopadhyay, F. Nowroozi, M. Ouellet, P. Peralta-Yahya, N. Prasad, S. Rodriguez, B. Rutherford, E. Steen, and J.D. Keasling

Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

The Fuels Synthesis Division of JBEI aims to engineer microbial host platforms and pathways for the production of fuels from lignocelluose hydrolysates obtained from JBEI's Deconstruction Division. Host platforms being employed by JBEI include *E. coli*, *S. cerevisiae*, and *Sulfolobus spp*. Host engineering includes a variety of tasks, such as metabolic engineering that combines native and non-native pathways, assessment of bottlenecks in metabolic flux and toxic effects of metabolites, gene discovery, and rational and combinatorial strain evolution. Collaborative efforts with JBEI's Technology Division include systems biology analytical and computational tools (genomics, transcriptomics, proteomics, metabolomics, fluxomics), robotics, and imaging. Of the broad range of fuels of interest to JBEI, efforts to date have focused on the following: (1) short-chain alcohols (e.g., butanol), (2) isoprenoid-based fuels (e.g., isopentenol), and (3) fatty acid-based fuels. In this poster, we present some of our strategies and results.

GTL

Building a *de novo* Synthetic Metabolic Pathway for Producing Branched-C5 Alcohols

H.H. Chou[1,3,4]* (HHChou@lbl.gov) and J.D. Keasling[1,2,3,4]

[1]Dept. of Bioengineering and [2]Dept. of Chemical Engineering, University of California, Berkeley, Calif.; [3]Joint Bioenergy Institute and [4]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofu-

els – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Many of the chemicals currently proposed as biofuels are derived from natural pathways, *i.e.* they already exist in nature. However, these natural chemicals were not designed to be fuels, so they have many costs associated with their use. The ability to build unnatural pathways to synthesize chemicals designed to have desirable fuel properties and be compatible with the current infrastructure would address many of the problems facing today's biofuels. This project demonstrates how pathways that don't exist in nature can be built, without the traditional approach of screening libraries of hundreds to thousands of enzymes, by taking advantage of the promiscuous nature of enzymes and enzyme superfamilies. Using this new approach, the project has constructed an unnatural pathway for synthesizing three different branched-C5 alcohols (3-methyl-3-butenol, 3-methyl-2-butenol, and 3-methyl-butanol) from the mevalonate pathway in *E. coli*. The three branched-C5 alcohols are promising biofuel candidates with many favorable fuel properties.

GTL

Transcriptomic Studies of the Response to Exogenous Exposure and Endogenous Production of Biofuel Candidates in *E. coli*

R. Dahl[1,3]* (robdahl@berkeley.edu), B. Rutherford,[1,3] A. Mukhopadhyay,[2,3] and J.D. Keasling[1,2,3,4]*

[1]Dept. of Chemical Engineering, University of California, Berkeley, Calif.; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [3]Joint BioEnergy Institute, Emeryville, Calif.; and [4]Dept. of Bioengineering, University of California, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

The production of biofuels in microbial systems presents a unique challenge to the host cell. Not only is the cell exposed to the solventogenic fuel molecule itself, the coordinated overexpression of an exogenous pathway presents a large burden on the cell's physiology, both in depleting nutrients and introducing foreign intermediates which could have toxic side-effects. At JBEI, we have focused on the impacts of short-chain alcohols (e.g. butanol) and isoprenoid-based fuels (e.g. isopentenol). Microarrays (in complementary studies with other system-wide "Omics" studies) were used to characterize *E. coli*'s response to these challenges. The general response has been one of a combination of oxidative, hyperosmotic (chaotrophic/ desiccation) and heat shock. In this poster, we present some of these results as well as the unique challenge to engineer the cell to better cope with these stresses.

GTL

Increasing Mevalonate Production by Engineering the Metabolism of *Escherichia Coli*

Yisheng (Connie) Kang* (srhconnie@gmail.com), Hector Garcia Martin, Peter Benke, Mario Ouellet, Jennifer Gin, and Jay D. Keasling

Joint BioEnergy Institute, U.S. DOE, University of California and Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Synthesis the anti-malarial drug, artemisinin, precursor mevalonate in *Escherichia coli* branches from acetyl Co-A which is the entry point to TCA cycle. We have been using genetic and environmental manipulations to redirect carbon flux from the endogenous central metabolic pathways (CMPs) to the heterologous pathway precursors. Deletion the production pathways of acetate, which is an undesirable product of excessive glycolytic flux, result in excretion of pyruvate rather than help increase mevalonate production. Heterologous mevalonate pathway from *Enterococcus faecalis* is more effective to draw carbon flow than that of *Saccharomyces cerevisiae*. Providing limited amount of nitrogen source also efficiently cut carbon flux to biomass and redirect it to mevalonate production. We are also performing metabolic flux analysis using 13C-labeled glucose. This information will help us determine how carbon flux through native metabolic pathways is affected by the presence of heterologous pathways, allowing us to identify and correct bottlenecks in the artemisinin production pathway. Successful completion of this project will provide insights into the metabolic status of living cells under different conditions and help us build a robust bacterium capable of producing high levels of artemisinin from cheap carbon sources.

GTL

Microbial Production of Isoprenoid Biodiesel

Taek Soon Lee* (tslee@lbl.gov), Nilu Prasad, Rossana Chan, Clem Fortman, and Jay Keasling

Joint BioEnergy Institute, Physical Bioscience Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Biodiesel, composed primarily of fatty acid methyl esters (FAMEs), has been one of the most successful renewable biofuels. They are derived from vegetable oil or animal fat. Limited supplies of feedstock, as well as competition for arable land, have been major drawbacks to expanding the use of biodiesel as alternative fuel. Microbial systems are used extensively for industrial small molecule production. Current techniques could be used to facilitate the production of biodiesel-like molecules in a microbial system. Furthermore, the agricultural industry generates a large and currently untapped supply of plant biomass (lignocellulose) that could be utilized as a renewable source of feedstock for these systems, resulting in carbon-neutral transportation fuels. In an effort to harness this potential, we have designed a new class of biodiesel fuels based on isoprenoids. We have chemically synthesized and tested a number of isoprenoid-based biodiesel candidates to determine which compounds are compatible with current engine technology. Biosynthetic pathways for the production of these validated isoprenoid-based fuel candidates have been designed in *E. coli*. Some of the pathways are constructed and tested for the production of target biofuels. The optimization of the production and the construction of more pathways genes are currently under investigation.

GTL

Harnessing Genomic Recombination to Improve Microbial Metabolic Phenotypes

Adrienne E. McKee* (AEMcKee@lbl.gov), John Haliburton, Veronica Fok, Mario Ouellette, Jay D. Keasling, and Swapnil Chhabra

Joint BioEnergy Institute, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

The microbial production of energy, pharmaceutical, and industrial compounds is a growing alternative to traditional, often costly, production processes. Many naturally occurring metabolic pathways of *Escherichia coli* and *Saccharomyces cerevisiae* have been enhanced for increased production of desired compounds. However, despite numerous advances, optimization of metabolic phenotypes faces many challenges as pathway improvement often requires both the redirection of intermediates and re-establishment of gene regulation – frequently requiring the modulation of many genes simultaneously. Furthermore, predicting the complement of genes that function cohesively for an organism to achieve a chosen metabolic phenotype may be exceedingly difficult, particularly if those gene products act at a distance from the pathway enzymes themselves. Genome shuffling (GS), a recently introduced strain improvement strategy, addresses these challenges through the use of genomic recombination to increase the genetic diversity of a population. When coupled with phenotypic screening and genome sequencing, GS holds the potential to discover genetic alterations that improve a phenotype as well as establish connections between gene products that may not otherwise be intuited from our current understanding of gene function or metabolic networks. Here, we present our recent efforts to develop protocols for protoplast fusion and genome shuffling in the industrial organisms *E. coli* and *S. cerevisiae*. We also discuss our current shuffling-based screens and selections that test the feasibility and effectiveness of GS in a directed application, namely increased production of the carotenoid 4,4′-diapolycopene, a product of the isoprenoid pathway. Through deep sequencing and comparative genomics, we will assess the new genotypes of strains that arise from this approach. As all isoprenoids share common metabolic precursors, the strains and genomic knowledge generated through this research may be applicable to the biosynthesis of a wide number of valuable industrial, pharmaceutical, and energy-related compounds. The results of this study will deepen our understanding of metabolic networks and will increase our knowledge of the diverse genomic landscapes that may converge on a select phenotype.

GTL

Optimizing Isoprenoid Biosynthesis

Farnaz Nowroozi* (FNowroozi@lbl.gov), David E. Garcia, Edward E. Baidoo, and Jay D. Keasling

Dept. of Bioengineering, University of California, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in

plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Isoprenoids are a diverse class of natural compounds that hold great commercial potential. The mevalonate pathway is one of the major pathways for isoprenoids biosynthesis. In the keasling lab, mevalonate pathway has been divided into two operons: the top pathway, and the lower pathway. We have shown that the lower mevalonate pathway is the limiting part of the mevalonate pathway and needs further optimization. We have shown that mevalonate kinase which is the first gene of the lower pathway is limiting. We are studying the kinetics of mevalonate kinase to determine the cause of this limitation. Also we are trying different ribosome binding sites with different strengths in front of the genes in the lower mevalonate pathway to optimize protein expression and therefore reach optimal metabolite production. For this purpose, it is desired to be able to measure pathway intermediates to determine which steps of the pathway are limiting, causing the intermediate metabolites to accumulate. We are trying to understand how the pathway expression is regulated when certain toxic intermediates accumulate.

We hope that this work would help us in further optimizing the isoprenoids biosynthetic pathway and constructing new pathways.

GTL

Metabolic Engineering of *Saccharomyces cerevisiae* for the Production of n-Butanol

Eric J. Steen[1,2*] (EJSteen@lbl.gov), Rossana Chan,[1,3] Nilu Prasad,[1,3] Samuel Myers,[1,3] Christopher J. Petzold,[1,3] Alyssa Redding,[1,3] Mario Ouellet,[1,3] and Jay D. Keasling[1,2,3,4]

[1]Joint BioEnergy Institute, Emeryville, Calif.; [2]Dept. of Bioengineering, University of California, Berkeley, Calif.; [3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [4]Dept. of Chemical Engineering, University of California, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Background - Increasing energy costs and environmental concerns have motivated engineering microbes for the production of "second generation" biofuels that have better properties than ethanol.

Results & Conclusions - *Saccharomyces cerevisiae* was engineered with an n-butanol biosynthetic pathway, in which isozymes from a number of different organisms (*S. cerevisiae*, *Escherichia coli*, *Clostridium beijerinckii*, and *Ralstonia eutropha*) were substituted for the Clostridial enzymes and their effect on n-butanol production was compared. By choosing the appropriate isozymes, we were able to improve production of n-butanol ten-fold to 2.5 mg/L. The most productive strains harbored the *C. beijerinckii* 3-hydroxybutyryl-CoA dehydrogenase, which uses NADH as a co-factor, rather than the *R. eutropha* isozyme, which uses NADPH, and the acetoacetyl-CoA transferase from *S. cerevisiae* or *E. coli* rather than that from *R. eutropha*. Surprisingly, expression of the genes encoding the butyryl-CoA dehydrogenase from *C. beijerinckii* (*bcd* and *etfAB*) did not improve butanol production significantly as previously reported in *E. coli.* Using metabolite analysis, we were able to determine which steps in the n-butanol biosynthetic pathway were the most problematic and ripe for future improvement.

GTL

Omics Research at the Joint BioEnergy Institute (JBEI)

Paul D. Adams, Tanveer S. Batth, Peter I. Benke, Rob Dahl, Jay D. Keasling, Aindrila Mukhopadhyay* (amukhopadhyay@lbl.gov), Mario Ouellet, Alyssa M. Redding, Becky J. Rutherford, Chris J. Petzold, and Pramila Tamrakar

Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Omics research at JBEI is part of the Technologies division and provides enabling tools for a variety of cell wide and analytical measurements required for a Biofuels research program. For the conversion of sugars derived from deconstructed lingo-cellulosic biomass to fuel compounds, an important area of research is on engineered organisms that contain combinations of native and non-native biochemical pathways for the production of a target metabolite and also efforts to understand the causes of toxicity/stress during such applications. For example, the incorporation of exogenous biochemical pathways into a host organism places unregulated strain on the cell by consuming metabolites, energy and critical cofactors creating an imbalance that will trigger a variety of stress response systems. While these stress responses may be useful to the cell in such an environment, they are unfavorable in an engineering context and reduce product yield or viability during production culturing. Systems biology, built on the foundations of omics studies (genomics, proteomics, metabolomics and fluxomics), enables a comprehensive view of the impact of an exogenous pathway on the host within the context

of its full metabolism. To match the requirement of such high-throughput profiling of particular classes of cellular components, genes, proteins and metabolites, our capabilities now include microarray analysis and high resolution mass spectrometry (combined with LC, GC, and CE for shotgun proteomics, targeted protein studies, primary and secondary metabolite analysis and glycomics). In collaboration with the computational core we also have powerful data analysis and integration tools. These functional genomics workflows are now being applied to gather data for the effect of (1) exposure of deconstruction conditions (Ionic liquids, post saccharification mix from simple and complex cellulose sources); (2) accumulation of endogenous and exogenous target metabolites; and (3) impact of different growth condition and expression of the biosynthetic pathway proteins in our model host microbes (*E. coli*, *S. cerevisiae*). Specifically the use of targeted proteomics using the MRM workflow has proved valuable in gauging the presence of a complete engineered pathway. Finally, the functional genomics and analytical tools described above are also being extensively used by the Feedstock division for a molecular characterization of cell wall and by the Deconstruction division in metagenomic studies.

GTL

JBEI Computational Biology Core

Dylan Chivian[1,2,3*] (DCChivian@lbl.gov), Paramvir S. Dehal,[2,3] Marcin P. Joachimiak,[2,3] Keith Keller,[2,4] Morgan N. Price,[2,3] Jason K. Baumohl,[2,3] John T. Bates,[1,3] and Adam P. Arkin[1,2,3,4]

[1]DOE Joint BioEnergy Institute, Emeryville, Calif.; [2]Virtual Institute for Microbial Stress and Survival, Berkeley, Calif.; [3]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [4]University of California, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Background: The Computational Biology Core group in the Technology Division of the Joint BioEnergy Institute (JBEI) is responsible for data integration and comparative, evolutionary, and functional genomic analysis for the purpose of enabling metabolic engineering for biofuel production. Leveraging the VIMSS MicrobesOnline website and database (http://www.microbesonline.org) for comparative and evolutionary genomics and analysis of microarray, proteomic, and metabolomic data sets, we are extending and integrating these capabilities to allow for pursuit of questions specific to biofuels challenges. These new tools will support the research of the Plant Feedstocks Division, the Deconstruction Division, and the Biofuels Synthesis Division of JBEI, and will be made available to the wider research community through the MicrobesOnline website.

Data Integration: The tools developed by the Computational Biology Core will interface with the functional data captured by a laboratory information management system (LIMS), including datasets from biomass production, growth curves, protein structures and imaging data, mass spec data, phenotype microarray data, regulatory motif and ChIP-chip data, transcriptomic, proteomic, metabolomic, metabolic flux, and genomic sequence data, as well as links to literature and other external databases. We are extending the abilities of MicrobesOnline to allow for integrated analyses of even more of the above data types, as such combined analyses of orthogonal data types allows for more complex questions to be asked and yields more robust results.

The MicrobesOnline Database: We have extended the MicrobesOnline database to include eukaryotic microbes that may be useful for understanding the process of biologically-mediated degradation of plant cell walls. Building a more complete picture of the enzymes nature employs for breaking down plant biomass is essential for developing industrial processes for biofuel production, and as such we are developing computational tools for metagenomics to permit searching for and comparative analysis of such enzymes in natural microbial communities. We are also working to combine computational structural biology with evolutionary analysis to grasp the mechanistic details of such "deconstruction" enzymes. This will permit prediction and engineering of novel enzymes with enhanced activities and custom specificities with an eye toward building a library of parts for metabolic engineering. Our efforts also include the study of the regulation of the expression and activity of cellulases and other deconstruction enzymes. Finally, we are extending the visualization and analysis tools in MicrobesOnline to provide a pathway-based view of systems to permit integrated analysis of systems biology data to facilitate metabolic engineering.

GTL

High Throughput Technologies to Break the Biological Barriers to Cellulosic Fuels

Masood Hadi[1,3*] (MZhadi@sandia.gov), H. Tran,[1,3] S. Morrison,[1,3] H.C. Chu,[1,3] A. DeGiovanni,[2,3] and P. Adams[2,3]

[1]Sandia National Laboratories, Livermore, Calif.; [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [3]Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

A recent paradigm shift in biological science research is from characterizing single genes or proteins (over an investigator's career) to studying whole genomes, proteomes or pathways in single "experiments" in a few months. It has been known for some time that while it is quite straightforward to clone and characterize a single gene. However, it is a completely different matter to perform detailed functional and structural studies in parallel for a few hundred genes, metabolic pathway or whole genome.

The ability to produce proteins is currently a major biological, physical, and computational challenge in protein research. Given a standard set of conditions, less than 30% of any given genome is expressible in a recombinant host. Protein expression requires complex, lengthy procedures, and specific proteins commonly require individual strategies for optimal expression. Standard bench-level procedures for protein production (expression and purification) do not exist. This lack of validated processes leads to a lengthy search for correct vector, host, expression and purification conditions to yield protein in milligram amounts. This problem is further compounded during metabolic engineering experiments where not only proteins have to be expressed, in addition genetic and regulatory processes have to be optimized for successful production of a product.

To this end the technologies division has been developing comprehensive suite of technologies in a consolidated facility. These successful methods and workflows are aimed at directed cloning methods for generating large numbers of expression constructs for protein expression and purification, screening of libraries for enzyme engineering and metabolic engineering. These processes will improve technical performance, productivity and reduce costs to allow affordability and timely progress towards our goals. We discuss some of these process development efforts and present initial results.

GTL

High Throughput Mass Spectrometry Based Enzymatic Assays for Biofuels Development

Trent R. Northen[1]* (trnorthen@lbl.gov), Jinq-Chyi Lee,[3] Chris Petzold,[2] Blake Simmons,[2] Paul Adams,[2] and Anup Singh[2]

[1]Life Sciences Division and [2]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, Calif.; and [3]Division of Biotechnology and Pharmaceutical Research, National Health Sciences Institute, Taipei, Taiwan

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Mass spectrometry's ability to efficiently generate intact biomolecular ions in the gas phase has led to a wide range of biological applications and is recently being applied for global metabolite profiling ('metabolomics') primarily though liquid chromatography coupled to electrospray mass spectrometry. However the complexity and relatively low throughput of this approach has limited application for high throughput enzymatic assays. To overcome this, we have developed the Nanostructure-Initiator Mass Spectrometry enzymatic (Nimzyme1) assay where enzyme substrates are immobilized on the mass spectrometry surface using fluorous phase interactions. This 'soft' immobilization allows efficient desorption/ionization while also allowing surface washing to reduce signal suppression from complex biological samples as a result of the preferential retention of the tagged products and reactants. We have also shown that Nimzyme can detect multiple and competing enzymatic activities and screen for optimal pH, temperature, and enzyme inhibition from crude cell lysates and a hot springs microbial community. This approach is being implemented at the DOE Joint BioEnergy Institute for high throughput functional characterization of both enzyme libraries and environmental samples. Specifically, we are constructing a complete set of glucose polysaccharides (cellobiose to cellihexose) for screening glucohydralase and glucotransferase activities and a p-coumaryl alcohol substrate for characterization of laccase activity. Together these assays will help to identify and optimize the conversion of lignocellulose into biofuels.

Reference

1. Northen, Trent R, Lee J-C, Hoang L., Raymond J., Hwang D-R, Yannone S.M., Wong C-H, Siuzdak G. A Nanostructure-Initiator Mass Spectrometry Based Enzyme Activity Assay, *PNAS*, (2008) 105(10), 3678-3683.

GTL

Multi-Mode Spectroscopic High Throughput Screening (HTS) of Phenols and Monolignols

Komandoor Achyuthan* (kachyut@sandia.gov), Blake Simmons, Paul Adams, and Anup Singh

Joint BioEnergy Institute (JBEI), Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Phenolic structure is a common motif among the monolignols, coniferyl alcohol, sinapyl alcohol and p-coumaryl alcohol. We are developing sensitive, rapid, multiplexed and high throughput screening (HTS)-compatible UV-Vis and fluorescence spectroscopic assays for phenols and

monolignols within 96- and 384-well microplates. We used p-cresol as a model for the phenols and coniferyl alcohol as a prototype monolignol. We employed the fungal (*Trametes versicolor*) laccase enzyme to oxidize p-cresol and coniferyl alcohol, and thereby expanded the spectroscopic properties of these molecules. Laccases are involved in lignin degradation. Our choice of laccase is especially relevant, since the enzyme was purified from white rot basidomycetes that are efficient degraders of lignins and widely studied for biofuels.

We supply a menu of spectroscopic options for the HTS of laccase oxidation of p-cresol through multiple modes of detection. Laccase activity was monitored kinetically at pH 4.5 by absorption changes at 250nm, 274nm or 297 nm, and in endpoint mode by the bathochromic shift in absorption to 326nm. Laccase oxidation of p-cresol was also detected by product fluorescence at 425nm after excitation at 262nm or 322nm. We optimized the kinetic parameters for p-cresol oxidation (pH optimum 4.5-5.1; 37oC; Km = 2.2mM) resulting in laccase limits of detection and quantization (LOD, LOQ) of 25pg/µL and 75pg/µL, respectively (~360pM; 25ppb). The p-cresol LOD was 8µM with a potential for further improvements in sensitivity. A key advantage of our assay is that laccase catalysis could be interrogated using multi-mode spectroscopy under acidic or basic conditions, in real time or endpoint modes.

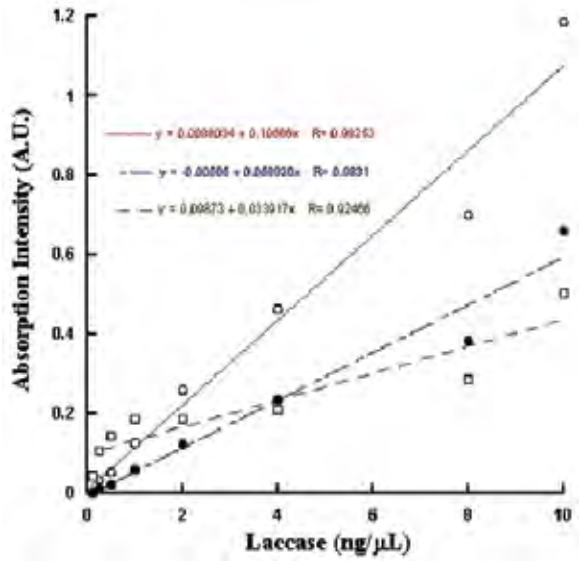


Figure 1. Absorption changes of laccase-catalyzed p-cresol oxidation. Laccase oxidation of 1mM p-cresol was in pH 4.5 buffer for 10 min. Absorption changes at 250nm (open circles), 297nm (closed circles) and 326nm (open squares) were measured as described above.

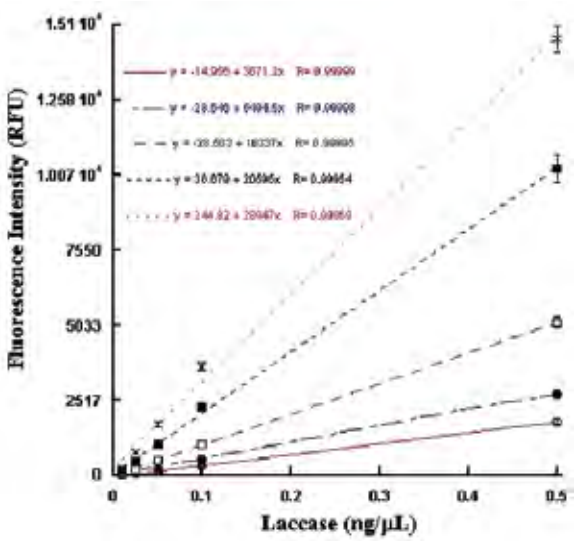


Figure 2. Laccase dose-response curves from 425nm fluorescence emission (excitation = 322nm). Reaction conditions were as described for Figure-1. Fluorescence was monitored for 2.5 min. (open circles), 5 min. (closed circles), 10 min. (open squares), 20 min. (closed squares) and 30 min. (crosses). All reactions were linear (r2 > 0.99).

We similarly characterized the spectroscopic properties of coniferyl alcohol in seven different solvents. Three isosbestic wavelengths were identified at 240nm, 242nm and 262nm between NaOH and the six solvents. A S/B of ~50 with 500µM coniferyl alcohol indicated assay sensitivity. The excitation spectrum was broad (270 – 335nm) and overlapped with absorption spectrum, as expected. Fluorescence emission was between 360 – 500nm with peak at 416 – 420nm. Fluorescence spectroscopy gave 1µM of coniferyl alcohol detection sensitivity. Unlike p-cresol, a fluorescence quench was observed following laccase oxidation of coniferyl alcohol.

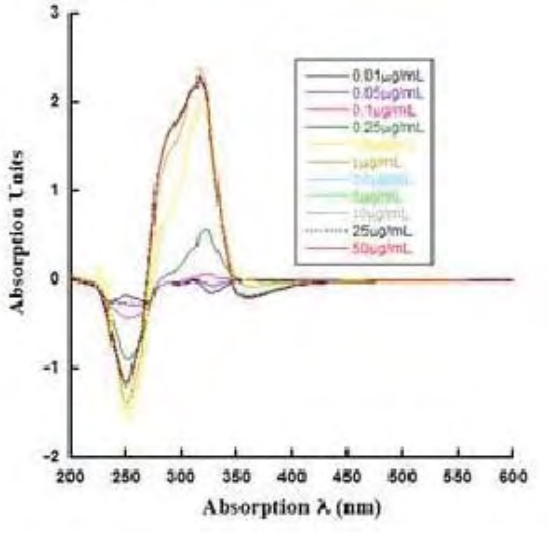


Figure 3. UV-Vis Difference Spectra of laccase-catalyzed oxidation of Coniferyl alcohol. Increasing concentrations of laccase were reacted in pH 4.5 assay buffer for 60 minutes with 1mM Coniferyl alcohol. The absorption changes taking place over the wavelength of 200 – 600nm are shown along with the concentrations of laccase used as inset.

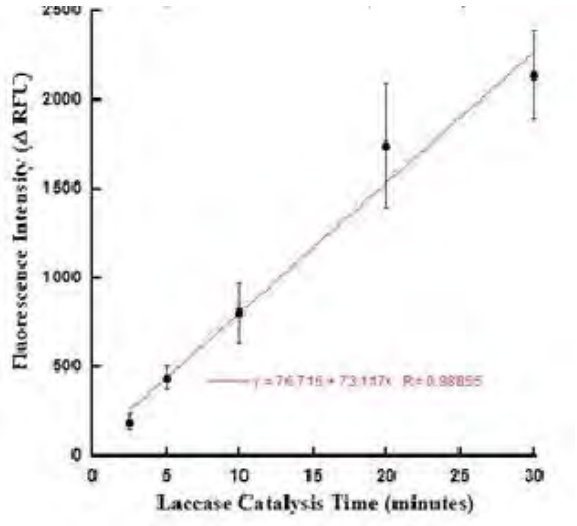


Figure 4. Laccase reaction kinetics from 416nm fluorescence emission (excitation = λ310 or 320nm). Reaction conditions were as described for Figure-1. Laccase activity was detected rapidly, in as little as 2.5 minutes towards the Coniferyl alcohol substrate.

In conclusion, we demonstrated sensitive, rapid, HTS-compatible fluorescence "turn on" and "turn off" spectroscopic assays for phenols and monolignols. Orthogonal interrogation and ratiometric analysis are key features of our assay, enabling high specificity and minimizing interferences during compound library screening. A portion of this work has been accepted for publication (below). We plan to expand our investigations to include the remaining two monoligonols: sinapyl and p-coumaryl alcohols.

Publication: K. E. Achyuthan, et al, Comb. Chem. HTS, 2009 (Accepted; In Press).

Acknowledgements:
This work was funded by the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under contract # DE-AC04-94AL85000.

GTL

Microfluidics for Protein Expression, Purification, and Screening

Rajiv Bharadwaj* (rbharadwaj@lbl.gov), Jim Brennan, and Anup K. Singh

Technology Division, Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

High-throughput (HTP) technologies for protein expression and screening are critical to the rapid development of cost-effective bioconversion processes. For example, methods are required for functional analysis of tens of thousands of wild-type and engineered lignocellulose degrading enzymes, as well as for full characterization of the plant cell wall synthesis machinery with a throughput of thousands of gene clones per year. We are developing novel integrated microfluidic platforms for small-scale (sub-µg levels) high-throughput protein expression, purification, and screening. The promising enzyme candidates can thus be rapidly identified and this paves the path for scaled-up production of enzymes.

Our microfluidic platform seeks to integrate fractionation of cell lysate, purification, characterization, and screening. A key to facile integration of these steps, which often features incompatibility of scale and modes of operation, is the replacement of centrifugation with liquid-liquid extraction for fractionation of cell lysates. We are also developing microfluidic cell-free-protein synthesis technology for simplified HTP protein expression. In addition, we are integrating on-chip protein characterization steps such as reversed-phase chromatography, SDS-PAGE, and native PAGE. Finally, we are developing colorometric and fluorescence-based enzyme assays for screening lignocellulases and glycosyltransferases.

GTL

Electron Microscopic Imaging at JBEI

Bernhard Knierim* (BKnierim@lbl.gov), Lina Prak, Seema Singh, Danielle Jorgens, Marcin Zemla, Kristen DeAngelis, Amitha Reddy, Jean Vandergheynst, Terry C. Hazen, Brad Holmes, Rajat Sapra, Blake Simmons, Paul Adams, and Manfred Auer

Joint BioEnergy Institute, Emeryville, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Advanced imaging approaches at JBEI include sophisticated sample preparation such as high-pressure freezing/freeze substitution, resin-embedded transmission electron microscopy (TEM), cryo-EM and electron tomography, as well as scanning electron microscopy (SEM). In addition we have advanced optical imaging techniques including confocal fluorescence microscopy, Raman microscopy, as well as atomic force microscopy. These techniques are applied to

* Presenting author

characterize feedstock cell walls, to determine subcellular protein localization in feedstocks, to monitor at high resolution the consequences on cell wall properties of ionic liquid pretreatment of biomass, to visualize microbial communities, and to analyze in detail the macromolecular lignocellulose degradation strategies of selected candidate microbes.

Specifically, we have examined the cell wall of *Brachypodium*, *Miscanthus*, *Equisetum* and found in 2D TEM projection views differences in density texture, most likely reflecting real differences in cell wall architecture.

Second, we have characterized the effect of pretreatment on cell walls as a function of length of exposure of ionic liquid pretreatment on Switchgrass biomass using high-resolution wide-field electron microscopy. We are planning to subject specimen from selected time points to electron tomographic 3D analysis.

Third, we have studied Puerto Rico rain forest and compost microbial communities and found an abundance of bacterial shapes and sizes, as well a variety of interesting extracellular features likely to be involved in lignocellulose degradation. Our images will complement phylogenetic profiling of such samples and may allow a spatial mapping of the respective position of community members and their interaction.

Fourth, we have studied *Sulfolobus* samples incubated in the presence of a variety of different substrates. Only in the presence of cellulose did we find an organelle-like feature that appears to be a novel membrane-bound cellulose-degrading extracellular specialization. We are currently in the process of characterizing this novel feature in more detail.

GTL

An MRM-based Mass Spectrometry Method for Optimization of Protein Expression to Increase Biofuel Production

Christopher J. Petzold[1]* (CJPetzold@lbl.gov), Alyssa M. Redding,[1] Sung Kuk Lee,[1] Farnaz F. Nowroozi,[2] Tanveer S. Batth,[1] Aindrila Mukhopadhyay,[1] Paul D. Adams,[1] and Jay D. Keasling[1,2]

[1]Joint BioEnergy Institute, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [2]Dept. of Chemical Engineering, University of California, Berkeley, Calif.

http://www.jbei.org

Project Goals: JBEI's primary scientific mission is to advance the development of the next generation of biofuels – liquid fuels derived from the solar energy stored in plant biomass. JBEI is one of three new U.S. Department of Energy (DOE) Bioenergy Research Centers (BRCs).

Optimization of biofuel pathway protein levels is crucial to balancing the energetic and carbon utilization of a microbe for efficient biofuel production. However, identification and quantification of specific proteins in complex mixtures is a difficult task. Since the physical attributes of proteins (e.g., MW, pI) are quite similar extensive separation or high specificity are needed to correctly identify a particular protein from a cell lysate. Western blots simplify analyses due to their high selectivity towards the target protein and tagging the protein of interest offer a means by which selective enrichment is possible. Yet, Western blots and tagging have limitations that make alternate methods attractive. One method, multiple-reaction monitoring, is capable of rapidly changing the target protein, something not possible without an antibody for the new protein, and detecting multiple target proteins in the same sample, something not possible without multiple tagging strategies and different enrichment steps. Multiple-reaction monitoring (MRM) is a mass spectrometric technique that has been used for small molecule DMPK studies for many years and has recently been adapted to peptides. Coupled to liquid chromatography, MRM-based analysis offers high selectivity and sensitivity. This method utilizes two points of selection (a peptide mass and a specific fragment mass generated by MS/MS) to eliminate background signal and noise even in very complex mixtures. Since the entire mass range is not scanned and only specific MRM transitions (combinations of peptide and fragment masses) are detected a significant increase in sensitivity is typically observed. Careful selection and optimization of MRM transitions permits detection of 5-10 specific proteins per LC-MS analysis. We are currently developing and optimizing MRM transitions to target proteins of interest for producing high titers of biofuel molecules. Our initial efforts are directed at optimizing the mevalonate pathway, the foundation for producing isoprenoid-based biofuels. The mevalonate pathway also serves as a good model for butanol-producing microbes. With these methods we will characterize a variety of protein expression parameters (promoter; ribosome binding site; plasmid) to determine optimal metabolite (e.g., mevalonate, isoprenoids, butanol) production.

Great Lakes Bioenergy Research Center (GLBRC)

GTL

Streamlined Method for Biomass Whole-Cell-Wall Structural Profiling

John Ralph[1*] (jralph@wisc.edu), Fachuang Lu,[2] Hoon Kim,[2] Xiaowen Fang,[2] Mark Norton,[2] James J. Ellinger,[2] John L. Markley,[2] Daniel Yelle,[3] Kenneth E. Hammel,[4] Sally A. Ralph,[4] Ronald D. Hatfield,[5] Paul F. Schatz,[5] Mattias Hedenström,[6] and Björn Sundberg[7]

[1]Dept. of Biochemistry (Enzyme Institute) and Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wis.; [2]Dept. of Biochemistry, University of Wisconsin, Madison, Wis.; [3]USDA Forest Products Laboratory and Dept. of Forestry, University of Wisconsin, Madison, Wis.; [4]USDA Forest Products Laboratory, Madison, Wis.; [5]U.S. Dairy Forage Research Center, Madison, Wis.; [6]Dept. of Chemistry and [7]Umeå Plant Science Centre, Umeå University, Umeå, Sweden

http://www.greatlakesbioenergy.org

Project Goals: The aim is to provide the plant cell wall and biomass research communities with improved methods for polysaccharide and lignin structural profiling, based on complete cell wall solubilization and NMR. The intent was to develop and streamline procedures to allow 20-30 samples per day to be profiled.

Introduction
In wide-ranging research aimed at altering plant cell wall characteristics, by conventional breeding or modern genetic methods, one of the biggest problems is in delineating the effects on the cell wall. Plant cell walls are a complex conglomerate of a variety of polysaccharides and lignin. Although other methods have their place, and can be more rapid (e.g. NIR), the difficulty in interpretation of some spectral methods, or the destruction of structure by chemical methods, assures that key features of cell walls benefiting, for example, biomass production and conversion are lost. A promising recent approach is the dissolution of the whole cell wall and high-resolution solution-state NMR analysis. These methods are providing promising approaches to detailing the compositional and (chemical) structural characteristics of the complex cell wall polymers that are the basis of biomass conversion efforts. A coordination of the developments from cell wall dissolution, high-resolution 2D NMR profiling, and chemometrics approaches to relate the NMR profile to other properties, seems poised to allow medium-throughput profiling of the wall, an area in which there are no currently comparable methods. The dissolution, NMR, and associated chemometrics methods are anticipated to impact research on optimal feedstock selection, aid in predicting biomass conversion efficiencies, and help process optimization.

Objectives
The aim was to provide the plant cell wall and biomass research communities with improved methods for polysaccharide and lignin structural profiling, based on complete cell wall solubilization and NMR. The intent was to develop and streamline procedures to allow 20-30 samples per day to be profiled.

Background to CW-dissolution/NMR/Chemometrics, and Progress
Traditionally, structural analysis of cell wall polymers has been via destructive analytical methods or via NMR on fractionated isolates. The problem with the latter is that isolation of components from the complex and interconnected wall polymers delivers them altered, in only low yields, and partitioned in ill-defined ways. Some years ago, Fachuang Lu developed two solvent systems that allowed full dissolution of finely-ground cell walls.[1] Acetylation of cell wall material dissolved by one of the solvents (DMSO-NMI) delivered acetylated walls that were soluble in common (organic) NMR solvents. The dispersive power of 2D (and even 3D[2]) NMR produced the most comprehensive structural profiling of all the polymers in the cell wall, without fractionation. Already it is finding utility in the elucidation of changes in polysaccharides and lignins in transgenic plants,[3-5] is overturning accepted theories on fungal rotting mechanisms [even extensively degraded material is soluble in these solvents],[6] has recently revealed new cell wall cross-linking mechanisms (Kim unpublished), and is beginning to be applied to other recalcitrant polymers.

In a logical development, Daniel Yelle prepared perdeutero-NMI allowing CW dissolution directly into DMSO-d_6/NMI-d_6 right in the NMR tube, without requiring the tedious isolation steps following acetylation.[7] Some structures resolved better, some worse, than when acetylated. A major advantage was that natural lignin and polysaccharide acetylation could be detected – acetylation is typically a detriment to efficient saccharification and therefore needs to be assessed. The drawback is that NMI-d_6 is expensive.

In attempts to find cheaper and simpler systems, Hoon Kim discovered that poorer quality but nevertheless impressive and informative spectra could be acquired by simply swelling the ball-milled cell walls in the readily accessible NMR solvent, DMSO-d_6.[8] More recently (Kim, manuscript in preparation), the addition of commercially available pyridine-d_5 was shown to improve the quality of the spectra to near DMSO-d_6/NMI-d_6 levels. Additionally, with selection of superior 2D ^{13}C–^{1}H correlation NMR pulse sequences and optimization of parameters, NMR throughput at the rate of 20-50 samples per day (compared to the original rate of one sample per day!) is now attainable;[8] longer acquisition times continue to result in superior spectra with better detail for minor components, so will still be required for some studies. With the use of modern cryogenically-cooled NMR probes, whole-cell-wall 2D NMR spectra can readily be acquired with just 10-30 mg of material.

Beyond the utility of these methods to analyze cell wall structure (identifying the effects of gene manipulation, for example) is the potential to utilize the 2D NMR cell wall

* Presenting author

profile itself directly in multivariate analysis. As this had not previously been accomplished with 2D NMR data of this type, we collaborated with colleagues at Umeå University (Mattias Hedenström, Björn Sundberg) to develop the methods necessary to convert the data and apply a range of chemometrics methods to them. Initial test studies using tension wood vs normal wood (poplar) and pectin methylesterase misregulated transgenics are beginning to attest to the power of multivariate methods on the NMR profile.[9-11]

Work continues on improving all aspects of the procedures. Although we have not yet succeeded, it is hoped to substantially improve the step that is now the bottleneck, ball-milling the cell wall, via a microfuge tube system. NMR methods continue to be improved, and a significant effort is on improved assignments, especially of the polysaccharide components. Chemometrics methods continue to evolve. The methodology is being applied to a wider array of samples. The big remaining aspect, beyond the initial objectives, is to provide automated assignments and quantification of components, and to more effectively database and search the increasingly massive NMR databases via interactions with the NMR Facility at Madison.

References

1. F. Lu and J. Ralph. Non-degradative dissolution and acetylation of ball-milled plant cell walls; high-resolution solution-state NMR. *The Plant Journal*, 2003, **35**, 535-544.
2. J. Ralph and F. Lu. Cryoprobe 3D NMR of acetylated ball-milled pine cell walls. *Organic and Biomolecular Chemistry*, 2004, 2, 2714-2715.
3. J. Ralph, H. Kim, F. Lu, J. H. Grabber, W. Boerjan, J.-C. Leplé, J. Berrio Sierra, M. Mir Derikvand, L. Jouanin and C. Lapierre. Identification of the structure and origin of a thioacidolysis marker compound for ferulic acid incorporation into angiosperm lignins (and an indicator for cinnamoyl-CoA reductase deficiency). *The Plant Journal*, 2008, **53**, 368-379.
4. A. Wagner, J. Ralph, T. Akiyama, H. Flint, L. Phillips, K. M. Torr, B. Nanayakkara and L. Te Kiri. Modifying lignin in conifers: The role of HCT during tracheary element formation in Pinus radiata *Proceedings of the National Academy of Sciences, USA*, 2007, **104**, 11856-11861.
5. J.-C. Leplé, R. Dauwe, *et al*. Downregulation of cinnamoyl coenzyme A reductase in poplar; multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell*, 2007, **19**, 3669-3691.
6. D. J. Yelle, J. Ralph, F. Lu and K. E. Hammel. Evidence for cleavage of lignin by a brown rot basidomycete. *Environmental Microbiology*, 2008, **10**, 1844-1849.
7. D. J. Yelle, J. Ralph and C. R. Frihart. Characterization of non-derivatized plant cell walls using high-resolution solution-state NMR spectroscopy. *Magnetic Resonance in Chemistry*, 2008, **46**, 508-517.
8. H. Kim, J. Ralph and T. Akiyama. Solution-state 2D NMR of Ball-milled Plant Cell Wall Gels in DMSO-d_6. *BioEnergy Research*, 2008, **1**, 56-66.
9. M. Hedenström, S. Wiklund, J. Ralph, U. Edlund and B. Sundberg. A novel approach for analysis of poplar wood using two-dimensional NMR spectroscopy on dissolved whole cell walls in combination with chemometrics *Physiologia Plantarum*, 2007, **130**, XIth Cell Wall Meeting, Abstract 22.
10. M. Hedenström, S. Wiklund, F. Lu, L. Gerbner, P. F. Schatz, B. Sundberg and J. Ralph. Analysis of plant cell walls using 2D NMR spectroscopy and multivariate analysis; Application to tension wood and PME-mis-regulated plants. *Plant Journal*, 2009, in preparation.
11. J. Ralph. High-Resolution Solution-state NMR of Unfractionated Plant Cell Walls; Potential for Biomass Selection and Process Optimization. in *2008 Pittsburgh Conference*, New Orleans, LA, Paper 1730-1735.

This research was supported by the Office of Science (BER), U.S. Department of Energy, and USDA-CSREES, Interagency Agreement No. DE-AI02-06ER64299.

GTL

Quantification of Whole Plant Cell Wall and Plant Metabolites Using Advanced 2D ^{1}H-^{13}C HMQC NMR Techniques

Xiaowen Fang,[1]* James Ellinger,[1] Mark Norton,[1] William M. Westler,[1] Michael Casler,[2] Shawn M. Kaeppler,[3] and **John M. Markley**[1] (markley@nmrfam.wisc.edu)

[1]National Magnetic Resonance Facility at Madison, Dept. of Biochemistry, University of Wisconsin, Madison, Wis.; [2]Dairy Forage Research Center, USDA-ARS, Madison, Wis.; and [3]Dept. of Agronomy, University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org
http://www.nmrfam.wisc.edu

Project Goals: Quantify the relative abundance of cell wall components such as specific lignin, cellulose and hemicellulose in various biomass sources.

In the conversion of lignocellulosic plant material to liquid bioethanol, the linkages between lignin and hemicellulose/cellulose represent an important research focus, because the presence of lignin severely blocks access of cellulases to the hemicellulose/cellulose energy components. Thus it is important to be able to quantify the relative abundance of cell wall components such as ferulate in various biomass sources. Multi-dimensional solution NMR spectroscopy of unfractionated cell wall material[1] dispersed in organic solvents[2] provides a powerful approach for screening for variations in wall structure. 2D ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC) spectroscopy has been applied widely in studies of biomolecules and natural products including cell walls.[1] However, the internal dynamics of solubilized cell wall polymers approach the rigid limit, and their shorter spin-spin relaxation times (T_2) and longer spin-lattice relaxation times (T_1) reduce signal sensitivity. We demonstrate here that these problems can be reduced through the use of sensitivity enhanced heteronuclear multiple quantum correlation (HMQC) spectroscopy. The use of gradient enhanced HMQC improves the uniformity of signal intensities across the spectral width and improves spectral dynamic range. This leads to improved quantification of signals from acetyl groups, whose presence correlates

with inhibition of saccharification. The improvements result in part from the longer characteristic T_2 of multiple quantum terms and the fewer number of radio frequency pulses required by the NMR experiment.[3] An additional benefit is that HMQC is more easily incorporated into pulse sequences designed for more detailed structural identification. In studies of switchgrass and gene mutated corn, have found that 2D ^{1}H-^{13}C HMQC data collection in combination with added internal standards allows accurate quantification of cell wall constituents. Ten samples per day can be analyzed in this way even without an automated NMR sample changer. Figure 1 illustrates the quantification of ferulate in switchgrass tissue from different species.

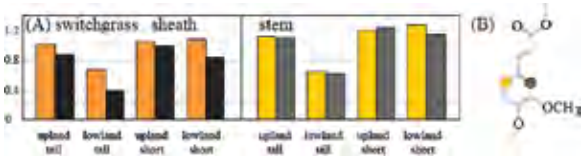


Figure 1. (A) Quantification of ferulate on the basis of 2D ^{1}H-^{13}C HMQC data in sheath and stem samples from upland tall and short and lowland of tall or short switchgrass. The orange and black bars represent data from two different CH groups of ferulate as shown in (B).

References

1. H. Kim, J. Ralph and T. Akiyama. Solution-state 2D NMR of ball-milled plant cell wall gels in DMSO-d_6. *BioEnergy Research.* 2008, **1**, 56-66.
2. D. J. Yelle, J. Ralph and C. R. Frihart. Characterization of nonderivatized plant cell walls using high-resolution solution-state nmr spectroscopy. *Magn Reson Chem*, 2008, **46**, 508-517
3. P. K. Mandal and A. Majumdar. A Comprehensive Discussion of HSQC and HMQC Pulse Sequences. *Concepts in Magnetic Resonance A,* 2004, **20A(1)** 1–23.

Acknowledgements
This work was carried out at the Great Lakes Bioenergy Research Center with support from the U. S. Department of Energy. We thank John Ralph and Hoon Kim for valuable advice of sample preparation.

GTL

Endogenous Variation for Biofuel Quantity and Quality Traits in Maize and Switchgrass

Shawn Kaeppler* (smkaeppl@wisc.edu), Natalia de Leon (ndeleongatti@wisc.edu), Mike Casler (mdcasler@wisc.edu), Heidi Kaeppler (hfkaeppl@wisc.edu), Rajan Sekhon, James Johnson, and Candy Hansey

University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Ethanol potential per unit of land is a function of biomass yield and biomass quality. Maize and switchgrass have substantial phenotypic variation among diverse cultivars, populations, and genetic collections for these traits. The goal of this research is to discover genes and alleles underlying this variation both to enhance the improvement of maize for biofuel production, as well as to use information from maize as a model for the closely related biofuel crop, switchgrass. Biomass yield and quality are being evaluated on genetic mapping populations and diverse inbred lines. In maize, we are studying nearly 6000 genotypes from 28 recombinant inbred line populations. These populations have a useful combination of balanced allele frequencies, known allelic composition, and have large phenotypic variation for biomass quantity and quality. These populations are used to identify quantitative trait loci underlying biofuel traits. To further define the specific genes and alleles underlying the quantitative trait loci, we are developing an association mapping population composed of more than 500 lines that will mature in our environment. These materials are being genotyped with anonymous SNP markers to allow us to account for population structure caused by pedigree relationships and geographic isolation. Sequence polymorphisms for candidate genes will then be scored and associated with observed phenotypic variation to identify causal genes and polymorphisms. Pathways that are identified in this process will be subject to further perturbation and analysis using molecular approaches such as proteomics and metabilomics and and genetic approaches including mutagenesis and transgenics. Genes identified in maize as important in controlling phenotypic variation for biofuel traits will then be evaluated in switchgrass using similar genetic materials and approaches. The first traits that we have measured are developmental traits related to biomass quality and quantity. Approaches to best evaluate quality are under investigation and currently include NIR prediction and wet laboratory analysis using methods developed to assess digestibility in ruminants as well as using high throughput assays to measure sugars, convertibility, and direct ethanol production.

GTL

Discovery of Genes that Mediate and Regulate Hemicellulose Biosynthesis

Kenneth Keegstra* (Keegstra@msu.edu), Markus Pauly (Paulymar@msu.edu), Curtis Wilkerson (Wilker13@msu.edu), David Cavalier, Jean-Christoph Cocuron, Jacob Jensen, Nick Thrower, and Yan Wang

Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Mich.

http://www.greatlakesbioenergy.org

Current knowledge regarding hemicellulose biosynthesis in plants is fragmented and incomplete. While some proteins, and the genes that encode them, have been identified, many gaps remain and little is known about how hemicellulose biosynthesis is regulated. One attractive strategy for identifying the required genes is to perform expression profiling during periods of rapid hemicellulose deposition. Many plants produce large quantities of specific polysaccharides

as storage polymers in developing seeds. Studies of these seed systems has been used to identify proteins involved in mannan and xyloglucan biosynthesis (1,2). We are extending these earlier studies by using 454 sequencing technology to perform deep EST sequencing at various stages of seed development during and just before rapid synthesis of mannan (Fenugreek-*Trigonella foenum-graecum*), xyloglucan (Nasturtium-*Tropaeolum majus*), or arabinoxylan (Psyllium-*Plantago ovata*). Analysis of the sequences obtained has confirmed the expression of genes known to be involved in the biosynthesis of these polysaccharides. In addition, a number of other genes have emerged as strong candidates for involvement in the production of these polysaccharides or in regulation of these pathways. Among the candidates that have been identified are putative sugar nucleotide biosynthetic enzymes, putative sugar nucleotide transporters, putative and known glycosyltransferases and glycansynthases of unknown specificity, proteins of unknown function, and transcription factors. Many of the genes have homologs that are expressed in developing wood or in other plant tissues where wall synthesis is occurring rapidly, providing support for the hypothesis that the same genes are involved in depositing these polymers in secondary cell walls. Promising examples from each class of candidate genes have been selected for detailed functional analysis. Selected examples from each polysaccharide will be presented on the poster.

References

1. Dhugga KS, Barreiro R, Whitten B, Stecca K, Hazebroek J, Randhawa GS, Dolan M, Kinney AJ, Tomes D, Nichols S, Anderson P (2004) Guar seed β-mannan synthase is a member of the cellulose synthase super gene family. *Science* 303: 363-366.
2. Cocuron J-C, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, and Wilkerson CG (2007). A gene from the cellulose synthase-like C family encodes a β-1,4 glucan synthase. *Proc Natl Acad Sci USA* 104: 8550-8555.

GTL

Biomass Trait Screening in a *Brachypodium* Mutant Population

Megan Galvin,[1] Sonali Nath,[1] Colleen Stephens,[1] Florian Kraemer,[2] Markus Pauly,[2] John Vogel,[3] and **John Sedbrook**[1]* (jcsedbr@ilstu.edu)

[1]School of Biological Sciences, Illinois State University, Normal, Ill.; [2]MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, Mich.; and [3]Western Regional Research Center, U.S. Dept. of Agriculture, Albany, Calif.

http://www.greatlakesbioenergy.org

The grass *Brachypodium distachyon* is emerging as an important model system for bioenergy crop grasses such as switchgrass and *Miscanthus* owing to its small genome size (~300Mbp), small stature, short generation time, transformability, and self-fertilization. The DOE Joint Genome Initiative (JGI) has recently sequenced the entire *Brachypodium* genome and is in the process of sequencing 300,000 *Brachypodium* Expressed Sequence Tags (ESTs). While others within the Great Lakes Bioenergy Research Center (GLBRC) are surveying natural variation of biomass traits within a collection of *Brachypodium* accessions collected worldwide, our group is focusing on identifying *Brachypodium* EMS mutagenized lines with differences in either enzymatic digestibility or stem morphology/growth characteristics. To date, using an HPLC based digestibility assay, we have screened through over 1,000 EMS lines and identified over 10 putative mutants with increased release of glucose and/or xylose upon hydrolytic enzyme digestion. We will discuss these results along with more detailed GC/MS-based cell wall analyses results. Our long-term goal is to identify genes and gene isoforms that can be introduced or bred into bioenergy crop grasses and corn thereby improving biomass digestibility/fermentability as well as biomass density and yields.

GTL

Understanding the Transcriptional Regulation of Secondary Wall Biosynthesis: A Step Toward Optimizing Lignocellulosic Feedstock for Biofuel Productivity and Processing

Jae-Heung Ko, Won-Chan Kim, Zhongnan Zhang, and **Kyung-Hwan Han*** (hanky@msu.edu)

Dept. of Forestry, Michigan State University, East Lansing, Mich.

http://www.greatlakesbioenergy.org

Secondary cell wall of plant biomass is a major source of lignocellulosic material, which is the most abundant natural compound on earth and a promising source of sugars for liquid biofuel production. Better understanding of the molecular mechanisms underlying its biosynthesis will help us develop biotechnological means to genetically control key pathways that determine the quantity and quality of the biomass. In an effort to identify candidate genes involved in transcriptional regulation of secondary wall biosynthesis, we developed an inducible secondary wall thickening system in *Arabidopsis thaliana*. Using this system, a battery of differentially expressed genes was identified. Most of the secondary wall biosynthetic genes (e.g., cellulose, hemi-cellulose and lignin genes) were induced within 6-hr of secondary wall thickening-induction treatment. We then identified several transcriptional regulators whose expression is coincided or preceded with the induction of secondary wall biosynthetic genes. The candidate genes, including C3H, LIM, NAC, MYB, and PXY transcription factors, are being tested in transgenic plants. Based on the transcriptome analysis, we constructed a tentative hierarchical transcriptional regulatory network leading to the biosynthesis of secondary wall components. In order to confirm the relationship between

transcription factors and their target genes, we are using both transient activation assay and electrophoretic mobility shift assay (EMSA). This poster presentation will describe (1) our genomics approach for identifying a transcriptional regulatory network that control secondary wall biosynthesis and; (2) functional characterization of selected candidate genes in the network.

GTL

Use of Proteomics Technologies for the Characterization of Proteins, Microbes and Microbial Communities Important for Bioenergy Production

Mary S. Lipton[1]* (mary.lipton@pnl.gov), Stephen J. Callister,[1] Kim K. Hixson,[2] Samuel O. Purvine,[2] Angela D. Norbeck,[1] Ronald J. Moore,[1] Therese R. Clauss,[1] and Carrie D. Nicora[1]

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

GLBRC Collaborators: Michigan State University: Shishir Chundawat, Dahai Gao, Venkatesh Balan, and Bruce Dale; University of Wisconsin, Madison: Robert Landick, Patricia Kiley, Tim Donohue, Eva Ziegelhoffer, Safak Yildaz, Daniel Noguera, Cameron R. Currie, and Jarrod J. Scott

http://www.greatlakesbioenergy.org

Microbial processing for substrates to biofuels, whether the conversion of ligno-cellulosic material to ethanol or microbial biorefineries to produce hydrogen or electricity is a central part of the GLBRC mission. Inherent in the use of microbes for these purposes is the characterization of the fundamental machines of the cell, the proteins, and how these proteins dictate microbial function. The quantitative determination of protein expression patterns and how these patterns change with changing cell state is critical for the GLBRC to remove bottlenecks in the bio fuels pipeline. Additionally, accurate measurements of protein levels and modifications will provide more extensive insights into both the plants and the microbes in the bioenergy pipeline. To this end, microbial researchers of the GLBRC will utilize state of the art proteomics facilities resident at the Pacific Northwest National Laboratory that allow for rapid global determination of protein expression patterns of cells or organelles. The proteomics facility at PNNL is one of these Genomics:GTL programs and as such will continue to characterize the protein complement of microbes and enzymes and is poised to make significant contributions in the characterization of microbial communities and plants.

Advances in proteomic technologies at PNNL have enabled characterization of unique features of microbial systems. Ranging from protein preparations purified from fungal and bacterial sources that demonstrate the ability to degrade lignocellulosic material to the quantitative proteomic profiling of microbes and microbial communities, the proteomics capability at PNNL is employed at producing data to further the understanding of systems important to the GLBRC. The proteomic analyses on all systems can be expanded to include temporal profiling through the analysis of time course studies, characterization of post-translational modifications, and determination of sub cellular localization of proteins. Extension of proteomic capabilities to community profiling will enable deeper understandings of how microbes interact with each other in environmental settings.

In the past year, the proteomics facility has supported the GLBRC in four aspects. The characterization of cellulolytic and hemicellulolytic rich enzyme cocktails found that the dominant cellulases were CBH I, Xyloglucanase, CBH II, EG I, EG II, EG III, β-glucosidase. The dominant hemicellulases found in most enzyme cocktails were Endoxylanase (GH 11), β- xylosidase, arabinofuranosidase (GH 62 & 54) & Glucuronidase (GH 67). This data will help determine critical classes of cellulases and hemicellulases necessary for hydrolyzing lignocellulosic biomass and are currently absent in commercially available mixtures. Supplementation of a minimalist and optimal enzyme set for hydrolyzing ammonia fiber expansion (AFEX) treated biomass (i.e. corn stover) will help reduce the total number and amount (mg protein/gm substrate) of enzymes required for hydrolysis.

The quantitative comparison of *E. coli* grown aerobically and anaerobically serves the preliminary basis for genetically engineering the organism to ferment ethanol. Analysis of aerobic and anaerobic cell cultures yielded a combined total of 1697 proteins identified from the two cultures using strict cross correlation and cutoff values for the peptides and the requirement of two peptides per proteins. Of the 1697 proteins identified, 46 proteins were found in the anaerobic cultures only, 30 peptides were found in the aerobic cultures only, and 1621 proteins were identified in both cultures. Quantitative analysis of 1254 proteins between both of the samples showed 86 proteins showed at least a two-fold increase in abundance in the aerobic sample and 56 showed at least a two-fold increase in abundance in the anaerobic sample. The qualitative and quantitative characterization of *Rhodobacter sphaeroides* proteome is being used to report on protein expression changes associated with solar- and feedstock-powered hydrogen accumulation.

Also, preliminary characterization of microbial communities isolated from Panamanian leaf-cutting ant colonies, which are known to degrade lingo-cellulose to organic carbon, show differences in microbial community protein expression between different sample areas of the nests.

GTL

Protein Expression Approaches to Cellulose Destruction

Nathaniel L. Elsen, Abolfazl Arabshahi, Allison Riederer, and **Brian G. Fox*** (bgfox@biochem.wisc.edu)

Dept. of Biochemistry, Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Project Goals: The goal of this project is to provide efficient methods for the discovery of new combinations of enzymes for improved destruction of cellulosic biomass.

Our long-term focus will be to provide a new combinatorial paradigm for evaluation of novel enzymes from new environmental sources as well as synthesized genes and engineered enzymes. The GLBRC bioenergy platform derives from work at the NIH Protein Structure Initiative-funded Center for Eukaroytic Structural Genomics, where over 10,000 genes from various eukaryotic organisms have been cloned, tested for expression, and in the best performing cases, purified and subjected to structure determinations. In this other project, over 1000 proteins have been purified, and over 100 protein structures have been determined. An adaption of the modular design of this platform provides the basis for this new effort on genes and proteins contributing to cellulose destruction. Vector design principles will be discussed, and a catalog of vectors available will be presented. Some of these vectors are already available in the public domain at the NIH Protein Structure Initiative Material Repository, http://www.hip.harvard.edu/PSIMR. Results of the application of our developed methods to characterization of the reactivity of various treated and untreated biomass materials will be presented.

This worked is funded by DOE DE-FC02-07ER64494 as a subgrant from the Great Lakes Bioenergy Research Center to B.G.F.

GTL

Engineering Cellulases with Improved Stability

George N. Phillips Jr.* (phillips@biochem.wisc.edu), Thomas J. Rutkoski, Robert W. Smith, and He Zhang

Great Lakes Bioenergy Research Center and Dept. of Biochemistry, University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Project Goals: To improve the efficiency or lower the cost of enzymes used to degrade cellulose to fermentable sugars for biofuel production.

Although significant progress has been made, the enzymatic hydrolysis of lignocellulosic feedstocks continues to be a significant factor affecting the economical production of cellulosic ethanol. Enzymes account for $0.10–0.25/gal of cellulosic ethanol produced. In it's 2006 publication, "Breaking Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda", the DOE has targeted a 10-fold reduction in cellulase cost as an R&D milestone. The goal of 1–2 cents for cellulase per gal of ethanol is comparable to the cost of amylase used in the production of ethanol from corn grain. Protein engineering efforts, incorporating both rational and directed evolution strategies, will be essential for reaching this ambitious goal. Specifically, enhancing the thermostability of these industrial enzymes would allow for higher specific activity, reduce the amount of enzyme loading during hydrolysis, and allow greater flexibility in process configurations. Here, we describe preliminary efforts to further enhance the conformational stability of cellulases.

Supported by the Department of Energy (DE-FC02-07ER644944).

GTL

Simple Chemical Transformation of Lignocellulosic Biomass into Fuels and Chemicals

Joseph B. Binder[1,2,*] and **Ronald T. Raines**[1,2,3] (rtraines@wisc.edu)

[1]Great Lakes Bioenergy Research Center, [2]Dept. of Chemistry, and [3]Dept. of Biochemistry, University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org
http://www2.biochem.wisc.edu/raines

Project Goals: Overcome biomass recalcitrance and transformation challenges through solvents tailored to biomass and innovative catalysis for carbohydrate conversion.

Lignocellulosic biomass is a plentiful, renewable, and largely untapped resource for fuels and chemicals. Despite this potential, nearly all renewable fuels and chemicals are now produced from edible resources, such as starch, sugars, and oils; the challenges imposed by notoriously recalcitrant and heterogeneous lignocellulosic feedstocks have made their production from non-food biomass inefficient and uneconomical.

Here, we report that *N,N*-dimethylacetamide (DMA) containing lithium chloride (LiCl) is a privileged solvent that enables the synthesis of the renewable platform chemical 5-hydroxymethylfurfural (HMF) in a single step and unprecedented yield from untreated lignocellulosic biomass, as well as from fructose, glucose, and cellulose (see Figure). Mechanistic analyses reveal that loosely ion-paired halide ions in DMA–LiCl are critical for the remarkable rapidity

(1–5 h) and yield (up to 92%) of this low-temperature (≤140 °C) process.

We also show that, in a second step, the HMF product can be converted into a promising liquid fuel: 2,5-dimethylfuran (DMF). DMF has an energy-content similar to that of gasoline and 40% greater than that of ethanol. Moreover, DMF is less volatile than ethanol and is immiscible with water. These attributes bode well for the use of DMF as an alternative fuel that is compatible with extant infrastructure.

Finally, we have demonstrated that our chemical approach enables efficient biomass saccharification to deliver sugars for fermentation processes.

Thus, a simple chemical transformation of lignocellulose can complement extant bioprocesses, providing a new paradigm for the use of biomass as a raw material for renewable energy and chemical industries.

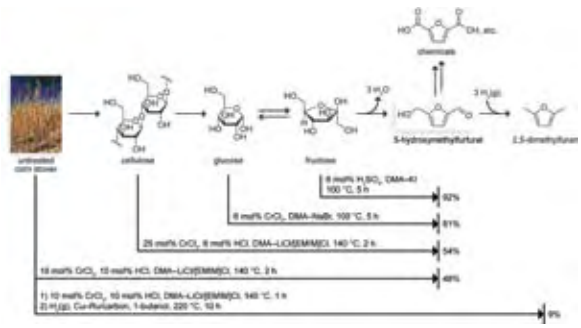


Figure. Halide salts in DMA enable previously elusive yields of bio-based chemicals from a variety of carbohydrates. Conditions are optimized for the conversion of carbohydrates into HMF (1 step) and DMF (2 steps). (Photograph courtesy of DOE/NREL.)

GTL

Construction of a Consolidated Bioprocessor Derived from *Escherichia coli*

David H. Keating* (dkeating@glbrc.wisc.edu), Mary Tremaine, and Robert Landick

Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Project Goals: The goal of GLBRC is to remove obstacles to economical microbial cellulosic ethanologenesis by generating new fundamental understanding of the cellular processes that underlie these bottlenecks.

Research at the Great Lakes Bioenergy Research Center aims to generate an improved understanding of the bottlenecks associated with the conversion of lignocellulose to ethanol. Our studies focus on the bacterium *Escherichia coli*, due to its sophisticated genetics, well-understood physiology, and use as an industrial microbe. We aim to construct consolidated bioprocessor *E. coli* strains, capable of the complete conversion of lignocellulose to ethanol.

The conversion of *E. coli* to a consolidated processor requires the introduction of heterologous genes responsible for cellulose degradation, as well as a secretory system for their transport from the cell. We are employing two parallel approaches to solve the secretion problem. First, we are introducing inducible promoters to activate the cryptic Type II secretory apparatus within *E. coli*. Second, we will engineer *E. coli* to express genes encoding the Type II secretory apparatus from closely related bacteria. We are also engineering *E. coli* to more efficiently produce and tolerate ethanol. To improve the ability of *E. coli* to produce ethanol from the C5 and C6 sugars generated from lignocellulose degradation, we are using candidate gene approaches, as well as random mutagenesis. Furthermore, we are employing metabolic modeling to identify novel combinations of mutations that link growth rate to the amount of ethanol production. Strains containing these combinations of mutations will then be subjected to directed evolution to identify variants with improved growth rates, with the expectation that such strains will also show improved rates of ethanologenesis. The mutants will then be subjected to resequencing, as well as a combination of metabolomics, proteomics, and transcriptomics, aimed at understanding the molecular mechanism behind the improved ethanologenesis. Collectively, these approaches will allow for the isolation of lead organisms that can then be subjected to further rounds of directed evolution. These studies will also produce an improved molecular understanding of the current limitations of ethanologenesis, and allow for the development of novel flexible approaches useful in diverse ethanologenic microorganisms.

GTL

Molecular, Genetic and Genomic Approaches to Alleviate Bottlenecks in Cellulosic Ethanol Production by Yeast

Trey Sato[1]* (tksato@glbrc.wisc.edu), Jennifer Van Vleet,[1,2,3] Thomas Jeffries,[2,3] Jeffrey Lewis,[1,3] and Audrey Gasch[1,3]

[1]Great Lakes Bioenergy Research Center, Madison, Wis.; [2]USDA Forest Products Laboratory, Madison, Wis.; and [3]University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Project Goals: The mission of the GLBRC is to perform the basic research needed to generate the technology that realizes the potential of cellulosic biomass to produce ethanol and other advanced biofuels.

Although progress has been made in engineering the yeast, *Saccharomyces cerevisiae*, to ferment cellulosic feedstocks, significant bottlenecks remain that limit the yield and efficiency of this process. These bottlenecks include

1) inhibition of fermentation by toxins generated in the pretreatment of cellulosic feedstocks and the fermentation process, 2) incapacity to hydrolyze cellulose and cellobiose into usable monomeric sugars and 3) an inability to ferment xylose, the second most prevalent monomeric sugar found in cellulose after glucose. The goal of our project is to utilize novel molecular, genetic and genomic approaches on the yeasts, *S. cerevisiae* and the xylose-fermenting *Pichia stipitis*, to develop a consolidated bioprocessing organism that can produce higher yields of cellulosic ethanol more efficiently than the current yeast standards. Our approach differs from other yeast engineering approaches in two key ways: 1) we will perform high-throughput screening of wild yeast strains to identify genetic backgrounds ideal for cellulosic fermentation, and 2) we will simultaneously engineer those strains for cellulase secretion, xylose fermentation, and multi-stress resistance, rather than optimize traits individually. This approach will advance the development of a single yeast strain with superior cellulosic fermentation properties, whereas optimization of single traits (e.g., xylose metabolism) often occurs at the expense of other desired phenotypes (such as stress tolerance). Here, we present our current strategies and ongoing research to engineer yeast at the Great Lakes Bioenergy Research Center for the production of ethanol from cellulosic feedstocks.

GTL

Catalytic Processing of Carbohydrates for the Production of Liquid Fuels

Drew J. Braden[1]* (braden@wisc.edu), Elif I. Gurbuz,[1] Edward L. Kunkes,[1] Dante A. Simonetti,[2] Ryan M. West,[1] Juan Carlos Serrano-Ruiz,[1] Christian A. Gartner,[1] and **James A. Dumesic[1]**

[1]University of Wisconsin, Madison, Wis.; and [2]University of California, Berkeley, Calif.

http://www.greatlakesbioenergy.org

Project Goals: Our goal is to develop a catalytic processing approach for the conversion of concentrated aqueous carbohydrate solutions in to liquid fuels. The catatlytic conversion approach involves the use of a combination of flow reactors operating in cascade mode where the effluent of one reactor is simply flown into the next. We have had success starting with feeds consisting of C_5 and C_6 sugars. Our goal now is to use feeds derived directly from lignocellulose. Specifically, we will examine the effects of solubilized lignin compounds on the catalyst surface chemistry. We will also investigate potential reactions for upgrading the solubilized lignin into liquid fuel components.

Biofuels can be produced in large volume from a wide variety of renewable sources through potentially carbon neutral processes [1]. In this respect, we have recently developed a catalytic approach for the conversion of aqueous C_6-sugar solutions, such as glucose and sorbitol, into gasoline, jet and diesel fuels [2]. The approach, shown in Figure 1, involves a combination of flow reactors operating in series, where the effluent of one reactor is simply fed into the next reactor.

Concentrated aqueous carbohydrate solutions (40-60 wt%) are initially reacted in a flow reactor over a Pt-Re catalyst. The carbohydrate species adsorbed on the catalyst surface undergo successive dehydration and hydrogenation reactions that effectively deoxygenate the molecule, increasing its hydrophobicity, as depicted in the bottom portion of Figure 1. In addition to the exothermic deoxygenation reactions occurring, part of the carbohydrate feed is converted to H_2 and CO_2 via endothermic reforming reactions. The *in situ* generation of H_2 eliminates the need for an external source of H_2 and balances the overall thermochemistry of the reaction, such that the overall conversion is mildly exothermic. The resulting reactor effluent consists of three phases: a gas phase, an organic phase and an aqueous phase.

The organic layer from the first catalytic processing step consists of monofunctional hydrocarbons in the form of alcohols, ketones, carboxylic acids and alkanes. This organic layer can be upgraded to components that are currently used in transportation fuels. For example, the monofunctional hydrocarbons can be combined via C-C coupling in aldol condensation reactions on $CuMg_{10}Al_7O_x$ to produce the C_8 to C_{12} linear alkanes used in diesel and jet fuel. Similarly, aromatic compounds used in gasoline can be produced from the organic species through aromatization reactions over an H-ZSM-5 catalyst. One advantage of this approach is that following the first processing step, more than 80% of the oxygen contained in the carbohydrate has been removed, allowing the subsequent upgrading processes to operate at reduced capacity and increased efficiency.

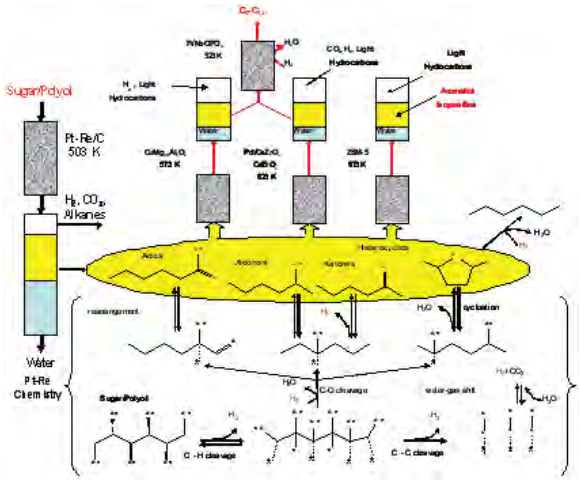


Figure 1. Schematic representation of carbohydrate processing for the production of monofunctional organic compounds, shown in yellow, and subsequent upgrading reactions in series. The proposed Pt-Re surface chemistries involved are shown in brackets where the asterisk signifies a catalyst site [2].

An essential requirement for overall effectiveness of our approach is to extend the process to utilize feeds derived directly from lignocellulose. Our initial work focused on the

C_6-sugar glucose, the monomer that makes up cellulose, as well as the C_6-polyol sorbitol. However, C_6 sugars only make up 25-40% of the total content of lignocellulose. Following our initial work, we have now successfully converted C_5 sugars, the main constituent of hemicellulose, to monofuntional organic intermediates with results comparable to C_6 sugar processing. With the combined processing of both C_5 and C_6 sugars, we are now able to convert approximately 70% of the content of lignocellulose into platform monofunctional organic compounds.

The next challenge in our work is to process feeds that include components that are not sugars, i.e., components derived from lignin. In particular, we are studying the effects of solubilized lignin components in mixed carbohydrate feeds to assess the role of these species in Pt-Re surface chemistry, and to determine whether these components can be upgraded to fuel components through our cascade mode catalytic approach.

References

1. B. Dale, *J. Agric. Chem.*, 56, 11 (2008).
2. E.L. Kunkes *et al., Science* 322, 417 (2008).

GTL

Engineering *E. coli* for Production of Hydrocarbons

Brian Pfleger (pfleger@engr.wisc.edu) and Rebecca Lennen* (lennen@wisc.edu)

University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Project Goals: We have developed a strain of *E. coli* that overproduces fatty acids through metabolic engineering. Our goals for this year include: discovering limiting biochemical steps through transcriptomic and proteomic analysis and metabolic modelling; further optimizing expression of acetyl-CoA carboxylase and acyl-ACP thioesterase; expressing and characterizing activity of a type I fatty acid synthase in *E. coli*; and implementing a high-throughput screen for discovery of improved fatty acid biosynthetic genes.

Long-chain hydrocarbons have higher energy densities than alcohols and are immiscible with water. They are therefore an attractive target for development of an engineered microorganism for sustainable fuel production. One potential platform for the synthesis of long-chain hydrocarbons in *Escherichia coli* is through reduction of fatty acids. The initial development of a fatty acid overproducing strain is reported. This strain features: (1) deletion of fadD, which encodes an acyl-CoA synthetase necessary for beta-oxidation; (2) overexpression of the four subunits of acetyl-CoA carboxylase (ACC), which converts acetyl-CoA to malonyl-CoA, a known bottleneck in fatty acid biosynthesis; and (3) heterologous expression of a codon-optimized plant medium chain acyl-acyl carrier protein thioesterase (BTE). Initial tests indicate an approximately ten-fold higher production of C_8 to C_{18} fatty acids in *E. coli* K-12 MG1655 Δ*fadD* when overexpressing ACC and BTE on plasmids. The predominant fatty acid chain length also shifts dramatically from C_{16} to C_{12} when BTE is expressed. Ongoing work includes optimization of ACC expression and identification of new metabolic and regulatory bottlenecks.

GTL

Networks Contributing to Photosynthetic Biohydrogen Production in *Rhodobacter sphaeroides*

Wayne S. Kontur[1]* (wskontur@wisc.edu), Eva C. Ziegelhoffer,[1] L. Safak Yilmaz,[2] Daniel R. Noguera[2] (noguera@engr.wisc.edu), and **Timothy J. Donohue**[1] (tdonohue@bact.wisc.edu)

[1]Dept. of Bacteriology and [2]Dept. of Civil and Environmental Engineering, University of Wisconsin, Madison, Wis.

http://www.greatlakesbioenergy.org

Project Goals: We seek to understand and optimize the light- and feedstock-powered production of hydrogen gas by the photosynthetic bacterium *Rhodobacter sphaeroides* for use as a clean and renewable source of energy. We are focusing our investigation on the various cellular pathways involved in hydrogen production. We want to determine the efficacy of hydrogen production from wild-type cells under different growth conditions, such as the carbon source used. We are analyzing these cultures via microarrays, proteomics, and other assays to determine the pathways that are utilized under the different conditions and, from this information, to gain insight into which pathways contribute to or detract from hydrogen production. We are also determining the effects of various mutations on hydrogen production. We will use these investigations to inform us on the design and development of strains, engineered through further mutation, that are optimized for hydrogen production under the conditions most relevant for large-scale systems.

Rhodobacter sphaeroides is a photosynthetic purple non-sulfur bacterium that can produce hydrogen gas from its nitrogenase enzyme, either concurrent with or independent of nitrogen fixation, during photoheterotrophic growth (and possibly from other, less-characterized sources). To assess the solar powered production of hydrogen under different conditions, *R. sphaeroides* wild-type strain 2.4.1 was grown photosynthetically in glutamate-containing medium (a poor nitrogen source that promotes the expression of nitrogenase) consisting of different carbon sources. Gas production analysis shows that the maximum rate of hydrogen production and the total amount of hydrogen produced from a culture depend on the identity of the carbon source used. We find general correlations between the ability of the bacteria to produce hydrogen and both the reducing potential

* Presenting author

and the chemical nature of the carbon source. Among the organic acids tested, the order of maximum production rate (in μL H_2/mL culture/hr) is: lactate (~90) > succinate (~61) > malate (~50) > tartrate (~32) > gluconate (non-detectable). The three sugars tested (glucose, xylose, and fructose) have similar maximum rates of hydrogen production, ~30-40 μL H_2/mL culture/hr. Cultures at the point of maximum hydrogen production rate were analyzed by microarrays, proteomics, and other assays to identify the metabolic pathways employed by *R. sphaeroides* that can contribute to or detract from hydrogen production. We find that a large percentage of the genes that are transcriptionally either up- or down-regulated in hydrogen producing cultures versus non-hydrogen producing cultures is related to metabolism. Among the genes up-regulated are those encoding subunits of the nitrogenase and hydrogenase proteins, as well as those related to electron transport, such as ferredoxins and flavins. Among the genes down-regulated are those involved in carbon fixation and the production of bioplastic polymers such as polyhydroxybutyrate, which are processes that compete with nitrogenase for reducing power. In addition to our studies of wild-type *R. sphaeroides*, we will report on studies to test the ability of various mutant strains (affected in electron flow pathways, light harvesting capacity, etc.) to produce hydrogen gas under comparable conditions.

GTL

Bacterial Communities in the Rhizosphere of Biofuel Crops as Evaluated by 16S rRNA Pyrosequencing

Ederson da Conceição Jesus* (ederson@msu.edu), Endang Susilawati (endangsusilawati@gmail.com), Qiong Wang (wangqion@msu.edu), Jorge Rodrigues (jorge@uta.edu) and **James M. Tiedje (**tiedjej@msu.edu)

Michigan State University, East Lansing, Mich.

http://www.greatlakesbioenergy.org

Project Goals: The goals of our project are to improve (a) the characteristics of biomass plants (Thrust 1); (b) the procedures for processing plant biomass (Thrust 2); (c) the biological or chemical processes to convert biomass into energy products (Thrust 3); and (d) the economic and environmental sustainability of the biomass–to–biofuel pipeline (Thrust 4). The specific goals of our group is to assess the structure and functional diversity of the rhizosphere community of biofuel energy crops, measure the extent of biological nitrogen fixation involved in biofuel crop production and describe and control interactions with symbiotic microbes and identify plant genes required for or regulating the establishment of AM symbioses.

The region around plant roots—the rhizosphere -harbors different microbial species which can fix nitrogen, protect plants against bacterial pathogens, produce plant growth factors and aid soil structure, thus providing for the overall promotion of plant growth. Managing these microbial communities can help to improve biomass production and decrease production costs as well. One of the steps toward reaching this goal is to know which species are present in the rhizosphere community and how environmental and plant factors affect microbial community structure. We studied bacterial communities in soils cultivated with several crops with potential to be used as biofuel crops, i.e. switchgrass, big blue stem, orchard grass, tall fescue, corn, soybean, canola and sunflower. Bulk and rhizosphere soil communities have been analyzed. 16S rRNA gene sequences have been amplified from community DNA with tagged-primers and sequenced by 454 technology, generating thousands of sequences that have been used for analysis of community structure and composition. Preliminary results from soils cultivated with switchgrass, big blue stem, orchard grass and tall fescue show that differences in bacterial community structure and composition were correlated to plant genotype and soil attributes, such as pH and nutrient concentrations. This correlation accounted for about 29% of the variation in the studied bacterial communities. Clearer differences were observed when rhizosphere soil was examined compared to bulk soil, which can be explained by the fact that the influence of plant genotype is stronger in the rhizosphere than in the bulk soil. Several phyla were found with *Proteobacteria*, *Acidobacteria* and *Actinobacteria* among the most abundant.

GTL

BioEnergy Cropping Systems on Marginal Land

Stephanie Smith* (smith579@msu.edu) and **Kurt Thelen** (thelenk3@msu.edu)

Michigan State University, East Lansing, Mich.

http://www.greatlakesbioenergy.org

Project Goals: Evaluate bio-energy crop yield and fuel quality to determine differences in crops grown on marginal land compared to agriculturally productive soils. Extrapolate results based on soil type and climate zone throughout Michigan to determine the optimal regional bio-energy cropping system.

Agriculture faces unique challenges as increasing world population places unprecedented demands on food and energy resources. Global fossil fuels are finite in supply and are becoming more expensive to extract as supplies are diminished. Bioenergy crops are increasingly being seen as essential components of future energy plans. There is particular interest in the production of these bio-energy crops on marginally productive lands. As agricultural resources are stretched to meet food and energy demands, lower productivity land bases will likely be brought into production.

Bioenergy Sciences Center (BESC)

GTL

The BioEnergy Science Center—An Overview

M. Keller,[1] L. Lynd,[2] A. Darvill,[3] and B. Davison[1]* (davisonbh@ornl.gov)

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Thayer School of Engineering, Dartmouth College, Hanover, N.H.; and [3]Complex Carbohydrate Research Center, University of Georgia, Athens, Ga.

http://bioenergycenter.org

Project Goals: The challenge of converting cellulosic biomass to sugars is the dominant obstacle to cost-effective production of biofuels in sustained quantities capable of impacting U.S. consumption of fossil transportation fuels. The BioEnergy Science Center (BESC) research program addresses this challenge with an unprecedented interdisciplinary effort focused on overcoming the recalcitrance of biomass. By combining engineered plant cell walls to reduce recalcitrance with new biocatalysts to improve deconstruction, BESC within five years will revolutionize the processing of biomass. These breakthroughs will be achieved with a systems biology approach and new high-throughput analytical and computational technologies to achieve (1) targeted modification of plant cell walls to reduce their recalcitrance (using *Populus* and switchgrass as high-impact bioenergy feedstocks), thereby decreasing or eliminating the need for costly chemical pretreatment; and (2) consolidated bioprocessing, which involves the use of a single microorganism or microbial consortium to overcome biomass recalcitrance through single-step conversion of biomass to biofuels. Within five years the Center will remove biomass recalcitrance as a barrier to cost-effective biofuels production by achieving a minimum two-fold reduction in the projected cost of processing for conversion of biomass to ethanol. Through this effort we will greatly enhance our understanding of cell wall structure during synthesis and conversion. The benefits of the basic research will extend beyond the five-year program by laying the foundation for developing other biomass sources and fuel products, improving productivity of switchgrass and poplar, and ensuring sustainability of lignocellulosic biofuel production.

In addition to Oak Ridge National Laboratory (ORNL), the BESC core team consists of the Dartmouth College, the University of Georgia, the Georgia Institute of Technology, the University of Tennessee, the National Renewable Energy Laboratory, the Samuel Roberts Noble Foundation, and industrial partners ArborGen, Diversa, and Mascoma. Other individual PIs complete the team. The home base of BESC is the Joint Institute for Biological Sciences building, funded by the state of Tennessee and occupied in December 2007. Located on the ORNL campus, the facility is designed specifically for interdisciplinary bioenergy research using systems biology tools. Other BESC anchor facilities include the University of Georgia's Complex Carbohydrate Research Center with extensive carbohydrate analytical and plant science expertise, the National Renewable Energy Laboratory's unique capabilities in comprehensive biomass analysis and bioprocessing, and ORNL's National Leadership Computing Facility

This poster will provide an overview of the BESC organization and year one activities. It will also provide context for the twenty other BESC posters which provide illustrative examples of specific BESC research efforts.

The BioEnergy Science Center is supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

The Use of TAIL PCR to Identify Genes Controlling Extreme Phenotypes in a *Populus* Activation Tagged Population

G.A. Tuskan[1]* (gtk@ornl.gov), T. Tschaplinski,[1] U. Kalluri,[1] T. Yin,[1] L. Gunter,[1] M. Martin,[1] N. Engle,[1] S. Jawdy,[1] X. Yang,[1] **Martin Keller**,[2] V. Busov,[3] S. Strauss,[4] X. Ma,[4] M. Davis,[5] and R. Sykes[5]

[1]Environmental Sciences Division and [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Michigan Technological University, Houghton, Mich.; [4]Oregon State University, Corvallis, Ore.; and [5]National Renewable Energy Laboratory, Golden, Colo.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC biomass formation and modification research involves working directly with two potential bioenergy crops (switchgrass and *Populus*) to develop varieties that are easier to break down into fermentable sugars. We are using both testing and generation of large number of natural and modified plant samples as well as detailed studies into poorly understood cell wall biosynthesis pathways. In the first year, the *Populus* activity has targeted testing of existing genetic resources (in association and activation tag studies). Direct transformation of *Populus* for over a hundred genes of interest has begun.

Bioenergy feedstock species provide the raw materials for biochemical conversion into sugars and ultimately liquid transportation fuels. The plant cell wall contains three major polymers including cellulose [i.e., a complex chain

of glucose molecules and the most abundant polymer on earth], hemicellulose [i.e., a complex mixture of five and six carbon sugars], and lignin [i.e., a polyphenolic matrix that protects the cell wall sugars]. Improving the yields of fermentable sugars from pretreated biomass is a key goal of the DOE Bioenergy Research Centers. Complete removal of lignin would be ideal for sugar yields but would be fatal to the plant. Thus, finding an optimum phenotype with a ratio of lignin to cellulose which results in high amounts of sugar *and* viability in the plant is one of our goals.

The detection of useful phenotypes is complicated by the long term nature of perennial woody plants such as *Populus*. Techniques to associate phenotypes with genotypes include conventional QTL studies, reverse genetics studies, naturally occurring mutants, as well as more sophisticated approaches such as activation tagging. During activation tagging a set of four tandemly repeated constitutive 35S promoters are randomly inserted into the genome of the target species. Plants are regenerated via tissue culture methods and the resulting intact plants are grown in the field for phenotypic evaluation.

Approximately 800 activation tagged lines of *Populus* were created in 2005 and established in field trials in eastern Oregon. Wood core samples and DNA were collected from these materials in February, 2008. Initial wood chemistry estimates, via molecular beam mass spectrometry (MBMS), were obtained from dried cores ground to 20 mesh size in a Wiley mill. Whole genome DNA template was extracted from leaf tissue and is being characterized via thermal asymmetric interlaced polymerase chain reaction (TAIL PCR). Sequences obtained from resulting TAIL PCR fragments are aligned to the *Populus* genome using BLASTN on the Joint Genome Institute (JGI) browser. Genes found within a ± 5 kb up and downstream region on the tag insert are candidate genes putatively affecting extreme wood chemistry phenotypes identified using MBMS.

Preliminary MBMS data suggests that the tested activation tagged lines have lignin values ranging from 17.6% (dry weight) to 23.0%, S/G ratios from 2.0 to 3.0, hemicellulose values from 20.0% to 25.6%, and cellulose values from 26.3% to 36.6%. Extreme phenotypes are defined as those values ± 2.5 standard deviations away from the wild type mean. Based on these criteria, 24 lines had at least one extreme phenotype, with three lines showing multiple extreme phenotypes. In these multiple phenotype lines it is likely that a single gene affected by the constitutive promoter is causing a change in the upstream portion of the carbohydrate biosynthesis pathway.

In order to identify the gene(s) associated with the extreme wood chemistry phenotype(s), TAIL PCR conditions were optimized using different combinations of reagent concentration, extension temperature and degenerative primers. Once conditions were optimized a defined tertiary reaction band was detectable on an agarose gel. These bands were excised, column purified and directly sequenced on an Applied Biosystem 3730 XL instrument. Individual reactions typically resulted in multiple bands, which may or may not be the same tagged locus [i.e. multiple insertions commonly occur during activation tagging]. BLASTN results on an initial set of tagged lines used for optimization indicate that unique positions within the *Populus* genome, and consequently the affected genes, can be identified using this technique.

The submitted abstract has been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science..

GTL

The pANIC Vector Set for Overexpression of Transgenes and RNAi-Mediated Knockdown of Native Genes in Switchgrass (*Panicum virgatum* L)

David G.J. Mann[1]* (dmann1@utk.edu), Peter R. LaFayette,[2] Laura L. Abercrombie,[1] Wayne A. Parrott,[2] C. Neal Stewart Jr.,[1] and **Martin Keller**[3]

[1]Dept. of Plant Sciences, University of Tennessee, Knoxville, Tenn.; [2]Center for Applied Genetic Technologies, University of Georgia, Athens, Ga.; and [3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC biomass formation and modification research involves working directly with two potential bioenergy crops (switchgrass and *Populus*) to develop varieties that are easier to break down into fermentable sugars. We are using both testing and generation of large number of natural and modified plant samples as well as detailed studies into poorly understood cell wall biosynthesis pathways. The Switchgrass activity is pursuing improved transformation methods, association studies and targeted pathways such as lignin synthesis.

Switchgrass (*Panicum virgatum* L) is a promising feedstock for the production of cellulosic ethanol, and switchgrass biotechnology plays a pivotal role in the BioEnergy Science Center initiative. A crucial component of creating transgenic switchgrass is having the capability of transforming the explants with DNA sequences of interest using vector constructs. However, there are very limited monocot plant vectors currently available. With this in mind, we have designed

and constructed a versatile set of 24 Gateway-compatible destination vectors (termed "pANIC") to be used in plants for transgenic crop improvement. Gateway compatibility allows for convenient insertion of any open reading frame (ORF) or other target sequence of interest. These vectors can be used for 1) transgene overexpression or 2) targeted gene silencing using double stranded RNA interference. Two main transformation methods exist for monocotyledonous plants: biolistic-mediated and *Agrobacterium*-mediated transformation. The pANIC vector set includes vectors which can be utilized for both applications, with all vectors containing three basic elements: 1) a Gateway cassette for overexpression or silencing of the target sequence, 2) a plant selection cassette and 3) a visual reporter cassette. The pANIC vector set allows for high throughput screening of sequences of interest in switchgrass, as well as other monocot plant species.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Genetic Modification of Lignin Biosynthesis in Switchgrass

Hiroshi Hisano,[1] Hui Shen,[2] Rangaraj Nandakumar,[1] Chunxiang Fu,[1] Xirong Xiao,[1] Yaxin Ge,[1] Tim Hernandez,[2] Fang Chen[2]* (fchen@noble.org), Richard A. Dixon,[2] Zeng-Yu Wang,[1] and **Martin Keller**[3]

[1]Forage Improvement Division and [2]Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, Okla.; and [3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC biomass formation and modification research involves working directly with two potential bioenergy crops (switchgrass and *Populus*) to develop varieties that are easier to break down into fermentable sugars. We are using both testing and generation of large number of natural and modified plant samples as well as detailed studies into poorly understood cell wall biosynthesis pathways. The Switchgrass activity is pursuing improved transformation methods, association studies and targeted pathways such as lignin synthesis.

Switchgrass (*Panicum virgatum*), a native C4 perennial grass throughout North America, is an excellent candidate for the production of cellulosic ethanol. The species has high biomass yield, low input requirements, good stress tolerance and the ability to grow in marginal areas. Switchgrass contains abundant sugars in the form of cellulose and hemicellulose, which cannot be easily converted to ethanol. On the other hand, Lignin, a major component of the cell wall of switchgrass, has been recognized for its negative impact on cellulosic ethanol production. Therefore, genetic modification of lignin in switchgrass could lead to increased ethanol production.

Our project focuses on genetic manipulation of key lignin genes to reduce recalcitrance to saccharification and to improve ethanol production in switchgrass. The project involves isolation of genes involved in lignin biosynthesis, construction of transformation vectors, production of transgenic switchgrass plants with modified lignin content/composition, and chemical analyses of the transgenic materials. In this presentation, we will show our target genes and strategy of transformation in switchgrass.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

New Insights on the Mechanism of Xylan Biosynthesis

Maria J. Peña[1]* (mpena@ccrc.uga.edu), Breeanna Urbanowicz,[1] Alan G. Darvill,[1,2] Malcolm O'Neill,[1] William S. York,[1,2] and **Martin Keller**[3]

[1]Complex Carbohydrate Research Center and [2]Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens, Ga.; and [3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC biomass formation and modification research involves working directly with two potential bioenergy crops (switchgrass and *Populus*) to develop varieties that are easier to break down into fermentable sugars. We are using both testing and generation of large number of natural and modified plant samples as well as detailed studies into poorly understood cell wall biosynthesis pathways. Biosynthesis research provides detailed studies in these and other plant models on key biosynthetic pathways for lignin, pectin, xylan and cellulose. The selection of target genes for transformation and validation has been coordinated and initiated.

Xylan accounts for up to 30% of the secondary wall where it is believed to interact with cellulose and lignin to generate biomass that is recalcitrant to deconstruction by microbial enzymes. Understanding how xylan is synthesized and incorporated into the wall may allow us to engineer plants with biomass that is more readily converted to biofuel.

Our studies of *Arabidopsis* mutants that form defective secondary cell walls (Peña et al., *Plant Cell 19:549-563* 2007) have shown that *Arabidopsis* xylan contains a unique glycosyl sequence at its reducing end that is required for normal xylan synthesis and that xylan structure and biosynthesis are more complex than previously believed. Hardwood and softwood xylans also have this sequence at their reducing ends. Thus, it may be possible to modify xylan properties in biomass by targeting genes that control the biosynthesis of this sequence.

We have developed an *in vitro* xylan biosynthesis assay using microsomal membranes, fluorescence-labeled oligosaccharide acceptors, and UDP-sugar donors and have structurally characterized the products formed. These studies show that *Arabidopsis* microsomes contain xylosyltransferase and glucuronosyltransferase activities that likely work together in the synthesis of glucuronoxylan. Switchgrass microsomes contain enzymes that in the presence of UDP-Xyl also extend the Xyl acceptors. Our results suggest that the xylan backbone is extended by a comparable mechanism in monocots and diocots. We present the results of these studies and discuss their implications for the mechanism of xylan biosynthesis.

Supported by U.S. Department of Energy, Bioenergy Research Centers, grant DE-PS02-06ER64304.

GTL

Functional Identification and Characterization of Sugar-1-P Kinases in *Arabidopsis*

Ting Yang[1,2,3]* (goldfishting@gmail.com), Liron Bar-Peled,[2] Lindsay Gebhart,[2] John Glushka,[2] Maor Bar-Peled,[1,2,4] and **Martin Keller**[5]

[1]BioEnergy Science Center, [2]Complex Carbohydrate Research Center, [3]Dept. of Biochemistry and Molecular Biology, [4]Dept. of Plant Science, University of Georgia, Athens, Ga.; and [5]Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC biomass formation and modification research involves working directly with two potential bioenergy crops (switchgrass and *Populus*) to develop varieties that are easier to break down into fermentable sugars. We are using both testing and generation of large number of natural and modified plant samples as well as detailed studies into poorly understood cell wall biosynthesis pathways. Biosynthesis research provides detailed studies in these and other plant models on key biosynthetic pathways for lignin, pectin, xylan and cellulose. The selection of target genes for transformation and validation has been coordinated and initiated.

The glycan salvage pathway is a metabolic process required during various stages of plant development: seed germination, pollen growth and embryo development, where storage carbohydrates are converted to simple sugars. The source of sugars for the salvage pathway likely utilizes monosaccharides derived from the hydrolysis of storage polysaccharides, free sugars that are released during wall remodeling, and during recycling of glycoprotein and glycolipids. Two major groups of enzymes exist in this pathway: sugar-1-P kinases and NDP-sugar pyrophosphorylases. Sugar-1-P kinases phosphorylate specific monosaccharide to their corresponding sugar-1-P, and subsequently the sugar-1-P may undergo pyrophosphorylation to form NDP-sugar.

D-galacturonic acid (D-GalA) is a major sugar residue found in pectin, plant primary cell wall polysaccharides, and in certain types of glycoproteins. Enzymes involved in synthesis of UDP-GalA in salvage pathway were identified over 40 years ago, yet the biological role of free GalA in plant metabolism remains elusive. Question remains as to the specificity of the kinase(s), and to the relative amount of GalA made in tissue; and the turn-over of pectin in various tissue to explain the need of GalA-kinase, remained unanswered. Here we report the discovery of a novel sugar-1-P kinase which phosphorylates GalA and to a lesser degree GlcA in *Arabidopsis*. The GalA-kinase was characterized and compared with galactose-kinase (GalK) using HPLC and NMR. To identify sugar and nucleotide specificities site-directed mutagenesis was applied.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

GAUT12 (GAlactUronosylTransferase 12): A Putative Glycosyltransferase Involved in *Arabidopsis* Secondary Cell Wall Biosynthesis

Zhangying Hao[1,3]* (haozhangying@gmail.com), Staffan Persson,[4] Kyle David Hutcheson,[2,3] Debra Mohnen,[2,3] and **Martin Keller**[5]

[1]Dept. of Plant Biology and [2]Dept. of Biochemistry and Molecular Biology, Complex Carbohydrate Research Center, BioEnergy Science Center (BESC), [3]University of Georgia, Athens, Ga.; [4]Max Planck Institute of Molecular

Plant Physiology, Potsdam, Germany; and [5]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC biomass formation and modification research involves working directly with two potential bioenergy crops (switchgrass and *Populus*) to develop varieties that are easier to break down into fermentable sugars. We are using both testing and generation of large number of natural and modified plant samples as well as detailed studies into poorly understood cell wall biosynthesis pathways. Biosynthesis research provides detailed studies in these and other plant models on key biosynthetic pathways for lignin, pectin, xylan and cellulose. The selection of target genes for transformation and validation has been coordinated and initiated.

GAUT12/IRX8 is a putative glycosyltransferase involved in *Arabidopsis* secondary cell wall biosynthesis. GAUT-type genes are widespread in plants (see poster by Y. Yin *et al*). *Arabidopsis* is being used as a model for functional studies to be confirmed when transformants in *Populus* and switchgrass are ready. The *irregular xylem 8* (*irx8*) mutant has reduced amounts of glucuronoxylan (GX) and, to a lesser degree, homogalacturonan (HG) in its walls. It was reported that the pentasaccharide sequence 4-β-D-Xyl*p*-(1-4)-β-D-Xyl*p*-(1-3)-α-L-Rha*p*-(1-2)-α-D-Gal*p*A-(1-4)-D-Xyl previously identified in birch (*Betula verrucosa*) and spruce (*Picea abies*) was also found at the reducing end of GX in *Arabidopsis*. Due to the sequence similarity of GAUT12/IRX8 to GAUT1, a known HG: GalAT, we propose two hypotheses for GAUT12 function. Hypothesis 1: GAUT12/IRX8 may be involved in the synthesis of the *Arabidopsis* GX: reducing end pentasaccharide by catalyzing the addition to the α-D-GalA onto the reducing end xylose. The moiety onto which the reducing end xylose is attached is not known. Hypothesis 2: The reduced amounts of GalA in *irx8* mutant walls suggests that GAUT12/IRX8 may be involved in the synthesis of a subfraction of HG (or possibly RG-I) to which xylan is covalently attached. Given the importance of understanding GAUT12/IRX8 function in wood formation and potentially in reducing biomass recalcitrance, we are working towards elucidating the enzyme function of GAUT12 by establishing robust biochemical assays for enzyme function. It is also our goal to identify potential protein partners that may interact with GAUT12 as it carries out its role in plant cell wall biosynthesis.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Isolation of Novel Biofuel-Relevant Thermophiles and the Identification of Extracellular Cellulolytic Enzymes Using Multi-Dimensional LC-MS/MS

James G. Elkins* (elkinsjg@ornl.gov), Scott D. Hamilton-Brehm, Adriane Lochner, Richard J. Giannone, Tatiana Vishnivetskaya, Jennifer Mosher, Mircea Podar, Jennifer Morrell-Falvey, Steve Allman, Dawn Klingeman, Sue Carroll, Miriam Land, Steven D. Brown, Robert Hettich, Tommy Phelps, and **Martin Keller**

BioEnergy Science Center, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC research in biomass deconstruction and conversion targets CBP by studying model organisms (e.g. *Clostridium thermocellum*) to understand novel strategies and enzyme complexes for biomass deconstruction. We are also searching natural diversity for enzymes and microbes—in particular, thermophilic anaerobes. Microbial research targets how the CBP microbe interacts with the features of the pretreated or native plant cell wall and testing strategies for using pure or mixed cultures for degradation.

Advanced conversion and fermentation technologies such as consolidated bioprocessing (CBP) offer the potential for a cellulose-based supply of liquid transportation fuels[1]. Ethanol production from CBP must include the use of robust microorganisms that are capable of efficient hydrolysis of lignocellulose with simultaneous fermentation of biomass-derived sugars to yield alcohol. Given the recalcitrance of biomass to enzymatic degradation, CBP must rely on a powerful suite of extracellular hydrolytic enzymes that are stable under what would likely be elevated temperatures. By integrating expertise in microbial ecology, microbiology, and omics-based approaches, we have established a microbial pipeline for the discovery and characterization of novel extremophiles that are able to break-down and metabolize plant-derived polymers.

Geothermally heated spring waters and sediments collected from Yellowstone National Park, Wyoming, USA, were selected as initial environments for screening and cultivation attempts. At total of 134 environmental samples were collected during two separate expeditions. Using these samples as a source of inoculum, enrichment cultures were established using dilute-acid pretreated switchgrass and *Populus*

* Presenting author

as the primary carbon and energy source. Temperatures for growth corresponded to the original sample temperature and ranged from 60-85°C. To isolate individual organisms from stable enrichment cultures, a high-throughput (HT) isolation system using flow cytometry was developed that allowed rapid separation and growth of cellulolytic, extreme thermophiles in liquid culture. Based on 16S ribosomal RNA gene sequences, isolated organisms clustered within the bacterial divisions Firmicutes and Dictyoglomi and several previously uncultivated species of *Caldicellulosiruptor, Thermoanaerobacter, and Dictyoglomus* were recovered. Secondary screening for rates of growth on pretreated biomass and crystalline cellulose showed that isolates from the *Caldicellulosiruptor* group possessed the fastest growth rates on relevant plant materials and model substrates. An organism isolated from Obsidian Pool, YNP, designated *Caldicellulosiruptor sp.* OB47 grows optimally at 80°C and reaches cell densities >10^8 cells/ml on carbon sources such as cellobiose, Avicel (crystalline cellulose), xylan, pectin, filter paper, processed cardboard, and pretreated lignocellulosic biomass (switchgrass and *Populus*). In batch growth experiments, OB47 produced end-products of >2.5 mM acetate, >50 mM CO_2, and >20 mM hydrogen. Comparative growth studies on 0.1% w/v Avicel between OB47 and known cellulolytic organisms, *Anaerocellum thermophilum*, *Caldicellulosiruptor saccharolyticus,* and *Clostridium thermocellum* revealed comparable growth rates which are the among the fastest known for cellulolytic organisms[2].

In order to gain more information regarding the functional enzyme system utilized by *Caldicellulosiruptor sp.* OB47 for growth on cellulose, genomic data was obtained from 454 pyrosequencing followed by assembly and annotation to construct a reference database of translated open reading frames. Initially, OB47 was grown on several cellulose-based substrates and cell pellets were harvested and prepared for proteomics analysis using multidimensional liquid chromatography mass spectroscopy (LC-MS/MS) to establish a baseline proteomic landscape. Using the translated draft OB47 genomic sequence as the reference protein database, roughly 1000 non-redundant protein identifications were obtained, corresponding to roughly 40% of the predicted expressed proteins. LC-MS/MS technology was then used to identify single components of the major extracellular hydrolytic enzymes expressed by OB47 during growth on filter paper. The extracellular protein fraction was also evaluated using filter paper cellulase assays to obtain general data about hydrolysis rate, optimum temperature, and enzymatic stability of the extracellular enzyme mixture. Over several biological and technical replicates, roughly 75 non-redundant proteins were identified by LC-MS/MS. Expectedly, several cellulases / glycoside hydrolases, cellobiosidases, solute-binding proteins, as well as S-layer domain-containing proteins were identified as the top hits when sorted by protein abundance. Using this analysis as a proof-of-concept, it was expanded to include growth on several different biofuel-relevant substrates such as switchgrass and *Populus*. In addition, this analysis can be readily expanded to include other candidate organisms which would provide a basis for comparison and allow for insights into different strategies (i.e. cellulosome vs. free enzymes) and particularly which families of glycosyl hydrolases individual microorganisms are expressing during growth on cellulosic materials.

References

1. Lynd, L.R., M.S. Laser, D. Bransby, B.E. Dale, B. Davison, R. Hamilton, M. Himmel, M. Keller, J.D. McMillan, J. Sheehan, and C.E. Wyman. How biotech can transform biofuels. *Nature Biotechnology* **26,** 169-172 (2008).
2. Lynd L.R., P.J. Weimer, W.H. van Zyl, and I.S. Pretorius. Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiol. Mol. Biol. Rev.* **66,** 506-577 (2002).

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Degradation of Plant Biomass without Pretreatment by the Thermophilic Anaerobe, *Anaerocellum thermophilum*

Irina Kataeva[1]* (kataeva@uga.edu), Sung-Jae Yang,[1] Nick Sennett,[1] Aleksandar Cvetkovic,[1] Farris L. Poole II,[1] Scott D. Hamilton-Brehm,[1,3] Richard J. Giannone,[3] Robert L. Hettich,[3] Nancy E. Engle,[3] Timothy J. Tschaplinski,[3] Janet Westpheling,[1,2] Michael W.W. Adams,[1,2] and **Martin Keller**[3]

Depts. of [1]Biochemistry & Molecular Biology and [2]Genetics, University of Georgia, Athens, Ga.; and [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC research in biomass deconstruction and conversion targets CBP by studying model organisms (e.g. *Clostridium thermocellum*) to understand novel strategies and enzyme complexes for biomass deconstruction. We are also searching natural diversity for enzymes and microbes—in particular, thermophilic anaerobes. Microbial research targets how the CBP microbe interacts with the features of the pretreated or native plant cell wall and testing strategies for using pure or mixed cultures for degradation.

We are investigating the conversion of switchgrass and the hardwood poplar by the anaerobic thermophile *Anaerocellum thermophilum*. Our goals are to understand at molecular level the mechanisms by which lignocellulosie biomass is degraded, and how these mechanisms differ with the type of plant biomass that is used and with defined substrates such as crystalline cellulose and xylan.

We have demonstrated that *Anaerocellum thermophilum,* an anaerobic and thermophilic bacterium that grows optimally at 75 ˚C, is able to degrade various types of unprocessed plant biomass, as well as defined carbohydrates such as crystalline cellulose and xylan. The plant materials that are utilized include hardwoods such as poplar, low lignin grasses such as napier and bermuda, and those with high lignin content such as switchgrass. The predominant reduced end products from all of these growth substrates are hydrogen and lactate. Glucose, cellobiose and, to a lesser extent, cellotriose (on crystalline cellulose) and xylose, xylobiose and some xylotriose (on xylan) also accumulated in the growth media during growth on the defined substrates, but these sugars did not accumulate during growth on the plant biomass. Substrate accessibility did not limit the growth of *A. thermophilum* on the plant material as the organism grew well on first- and second spent biomass derived from poplar and switchgrass. Spent biomass is the insoluble growth substrate recovered after the organism has reached late stationary phase. Extracellular extracts of *A. thermophilum* grown on various types of biomass have been subjected to proteomic, metabolomic, chromatographic, metal and enzymatic analyses. The results are being used to provide insights into the mechanisms of plant biomass conversions. We have also shown that a close relative of *A. thermophilum*, *Caldicellulosiruptor saccharolyticus* (T_{opt} 70˚C), grows well on switchgrass but not on poplar. Additional insights into mechanisms of biomass conversion are being obtained from a comparison of the genome sequences of these two thermophilic anaerobes.

This research was funded by the BioEnergy Science Center, which is supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Integration of Genomics and Bioinformatics to Identify Genetic Differences in an Ethanol Tolerant *Clostridium thermocellum* ATCC27405 Strain

Steven D. Brown[1]* (brownsd@ornl.gov), Tatiana Karpinets,[1] Jonathan R. Mielenz,[1] Shihui Yang,[1] Dawn M. Klingeman,[1] Miriam L. Land,[1] Loren J. Hauser,[1] Babu Raman,[1] Miguel Rodriguez Jr.,[1] Tingfen Yan,[1] Tatiana A. Vishnivetskaya,[1] Herbert J. Strobel,[2] Maria V. Sizova,[3] Nicolai Panikov,[3] Ying Xu,[4] Phuongan Dam,[4] Lee Lynd,[3] and **Martin Keller**[1]

[1]Biosciences Division and BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Dept. of Animal and Food Sciences, University of Kentucky, Lexington, Ky.; [3]Thayer School of Engineering, Dartmouth College, Hanover, N.H.; and [4]Institute of Bioinformatics, University of Georgia, Athens, Ga.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC research in biomass deconstruction and conversion targets CBP by studying model organisms (e.g. *Clostridium thermocellum*) to understand novel strategies and enzyme complexes for biomass deconstruction. We are also searching natural diversity for enzymes and microbes—in particular, thermophilic anaerobes. Microbial research targets how the CBP microbe interacts with the features of the pretreated or native plant cell wall and testing strategies for using pure or mixed cultures for degradation.

Clostridium thermocellum is a gram-positive, anaerobic, thermophilic bacterium that can ferment cellulose at one of the highest growth rates directly to ethanol via a large extracellular enzyme complex termed the cellulosome. *C. thermocellum* is a candidate industrial biocatalyst for future lignocellulosic fuel production. The metabolic byproducts of fermentation can inhibit fermentation performance and lignocellulosic biomass pretreatment processes also produce a variety of inhibitory chemicals that can adversely affect the fermentation. Limited information is available on the mechanisms and responses of *C. thermocellum* to different inhibitors. The genetic differences between wild-type *C. thermocellum* and an ethanol tolerant mutant have been identified through microarray based comparative genome sequencing and 454-pyrosequencing. We detected more than 400 differences in the ethanol tolerant mutant compared to the *C. thermocellum* wild-type strain. The resequencing data were in agreement with published membrane proteomic data and identified new mutations in key genes such as alcohol dehydrogenase. Bioinformatics analyses identified 16 mutational hot-spots in the ethanol tolerant strain, with 7 out of 16 related to cellulose degradation and likely accounted for the strain's decreased growth on cellulose. Further work to identify and verify important loci and physiological changes conferring tolerance to inhibitors will assist in the development of industrial strains for consolidated bioprocessing (CBP) of lignocellulosic biomass and therefore reduce biofuel production costs.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Advances in Microbial Cellulose Utilization: Methods Development and Kinetics

Nicolai Panikov[1]* (Nicolai.Panikov@Dartmouth.edu), Maria Sizova,[1] Evert Holwerda,[1] Julie Paye,[1] Lois St. Brice,[1] Lee Lynd,[1] and **Martin Keller**[2]

[1]Thayer School of Engineering, Dartmouth College, Hanover, N.H.; and [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC research in biomass deconstruction and conversion targets CBP by studying model organisms (e.g. *Clostridium thermocellum*) to understand novel strategies and enzyme complexes for biomass deconstruction. We are also searching natural diversity for enzymes and microbes—in particular, thermophilic anaerobes. Microbial research targets how the CBP microbe interacts with the features of the pretreated or native plant cell wall and testing strategies for using pure or mixed cultures for degradation.

Microbial cellulose utilization has great biotechnological promise, particularly for developing microbes capable of mediating consolidated bioprocessing. However, current understanding is profoundly limited with respect to most aspects of microbial cellulose fermentation. Methods for accomplishing tasks such as substrate and cell quantification are cumbersome, and the literature contains few if any controlled quantitative comparisons of the rates of microbial cellulose utilization. Owing partly to the absence of such comparative studies, we have no idea under what conditions microbial cellulose utilization occurs fastest. The first mechanistic kinetic model for microbial cellulose utilization has yet to be proposed.

BESC-supported research underway at Dartmouth aimed at meeting these research needs will be summarized. Particular topics addressed include:

- Development of advanced techniques for monitoring fermentation of insoluble cellulose, including inferring the cellulose concentration based on on-line measurements;
- Development of a quantitative assay for measuring and comparing rates of microbial cellulose utilization, functionally similar to the filter paper assay used for enzymatic studies;
- Development of a quantitative proteomic assay for the total cellulase concentration, including free cellulase enzymes as well as cell-associated and cellulose-associated enzymes; and
- Documentation and investigation of an apparent "fast growing" phenotypic variant of *Clostridium thermocellum* able to utilize cellulose substantially more quickly than strains described in the literature.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Metagenomics for Mining New Deconstructive Enzymes, Exploring Enzyme Diversity and Screening Cellulolytic Activities

Luen-Luen Li,[1] Sean M. McCorkle,[1] Denise C. Monteleone,[1] Susannah G. Tringe,[2] Tanja Woyke,[2] Shi-You Ding,[3] Michael Himmel,[3] Safiyh Taghavi,[1] Carl Abulencia,[4] Deborah Balch,[4] Ying Hefner,[4] Melisa Low,[4] Steven Truong,[4] Peter Luginbühl,[4] Steve Wells,[4] Joel Kreps,[4] Kevin Gray,[4] Daniel van der Lelie[1]* (vdlelied@bnl.gov), and **Martin Keller**[5]

[1]Brookhaven National Laboratory, Upton, N.Y.; [2]DOE Joint Genome Institute, Walnut Creek, Calif.; [3]National Renewable Energy Laboratory, Golden, Colo.; [4]Verenium Corporation, San Diego, Calif.; and [5]Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC research in biomass deconstruction and conversion targets CBP by studying model organisms (e.g. *Clostridium thermocellum*) to understand novel strategies and enzyme complexes for biomass deconstruction. We are also searching natural diversity for enzymes and microbes—in particular, thermophilic anaerobes. BESC is studying and modeling the structure and activities of multi-enzyme complexes to design or screen for variants with better cell wall deconstruction capabilities.

Plant biomass is the most abundant biopolymer on earth and has long been recognized as a potential sustainable source of mixed sugars for bioenergy production. Our goals are to understand the diversity, structure, functional interdependence, and metabolic capabilities of the natural cellulolytic microbial assemblages, and to exploit their dynamics for the conversion of plant biomass to a useful feedstock for biofuels production. Using metagenomics as an approach

allows the discovery of new enzyme diversity from microbial communities, especially from organisms that are unknown or have never been cultivated. From a microbial community actively decaying poplar biomass under anaerobic conditions, metagenomic DNA was isolated for further investigation. The distribution of microbial species in the community was investigated via 16S and 18S rRNA genes sequencing. *Saccharomycetes* composed the major group among the Eukaryotes, and *Clostridiales* composed the major group among the Bacteria. No major population of Archaea was found as part of the microbial community. Using the 454 GS FLX Titanium pyrosequencing, approximately 580 Mbp metagenomic DNA was sequenced. Preliminary homology searches of metagenome sequences revealed a high diversity of glycosyl hydrolase homologs (approximately 4,000 glycosyl hydrolases were identified). Five candidate glycosyl hydrolases were initially selected for further investigation, based on homology to enzyme families of interest (GHase families 5, 9, 48, and 51 representing cellulase, hemicellulase and xylanase activities) and the quality of the sequences (length, homology, potential gene rearrangements, disruptions, deletions). Full-length open reading frames of these genes were obtained by using inverse PCR and DNA walking, and gene cloning is presently in the process. Another approach to discover new glycosyl hydrolases is by constructing lambda-based expression libraries and screening clones for glycosyl hydrolase activity. Libraries were constructed from a variety of likely cellulolytic environments such as the digestive tract of herbivorous mammals and insects, microbial 'biotraps' and the gills of marine shipworms. These libraries will be screened using Verenium's ultra high-throughput GigaMatrix® system that can screen up to 1 billion samples per day. In a preliminary screen, 10 million clones were screened and 353 primary hits identified from two environmental libraries. The number of active clones reduced to 61 after the tertiary screen and based on DNA sequence data, 29 unique, active enzymes were identified, 14 of which have known GH domains. Additional activity screening will be performed on these libraries, as well as large-scale sequencing at JGI of the original environmental DNAs. Our combined metagenomic studies and enzyme activity screens will provide insight into the microbial community compositions as well as provide a resource for discovering diverse, novel, community-encoded glycosyl hydrolases.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

The Improved Cellulosome: Computational Modeling to Minisomes

Yannick J. Bomble[1]* (yannick_bomble@nrel.gov), Michael F. Crowley,[1] Shi-you Ding,[1] Qi Xu,[1] Mark R. Nimlos,[2] Jiancong Xu,[3] Moumita Saharay,[3] Hong Guo,[4] John W. Brady,[5] David B. Wilson,[6] Jeremy C. Smith,[3] Michael E. Himmel,[1] and **Martin Keller**[7]

[1]Chemical and Biosciences Center and [2]National Bioenergy Center, National Renewable Energy Laboratory, Golden, Colo.; [3]Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [4]Dept. of Biochemistry and Cell and Molecular Biology, University of Tennessee, Knoxville, Tenn.; [5]Dept. of Food Science and [6]Section of the Biochemistry, Molecular, and Cell Biology, Cornell University, Ithaca, N.Y.; and [7]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC research in biomass deconstruction and conversion targets CBP by studying model organisms (e.g. *Clostridium thermocellum*) to understand novel strategies and enzyme complexes for biomass deconstruction. We are also searching natural diversity for enzymes and microbes—in particular, thermophilic anaerobes. BESC is studying and modeling the structure and activities of multi-enzyme complexes to design or screen for variants with better cell wall deconstruction capabilities.

This study aims at understanding the mechanisms involved in the sequential binding of the cellulosomal enzymes on the scaffold of *C. thermocellum*. The first coarse grained model to study the formation and function of the cellulosome was developed within CHARMM. Some of the binding constants between cohesins and dockerins were derived from all-atom simulations. Individual subdomains were also studied with CHARMM and Amber on cellulose surfaces or with individual cellulose chains. These domains include catalytic domains, carbohydrate binding domains, and fibronectins. All five cellulosomal fibronectins (Fn3) of *C. thermocellum* have been identified, overexpressed and purified, the crystal structure of one of them has been solved, and this provided experimental structure for the computational modeling of cellulosomal Fn3 function.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

* Presenting author

GTL

A Rapid Analytical Pyrolysis Method for Investigating Genetic Modification of the Lignin Pathway in Alfalfa (*Medicago sativa*)

A. Ziebell[1]* (angela_ziebell@nrel.gov), R. Sykes,[1] D. Astling,[1] K. Reichel,[1] C. Doeppke,[1] M. Davis,[1] and **Martin Keller[2]**

[1]National Renewable Energy Laboratory, Golden, Colo.; and [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Knowledge gained by thoroughly characterizing biomass chemistry and structure will drive coordinated development of improved plant biomass and degradation microbes. Native, genetically modified, and partially deconstructed lignocellulosic samples will be analyzed.

Pyrolysis-Molecular Beam Mass Spectrometry (py-MBMS) and pyrolysis-Gas Chromatography Mass Spectroscopy (py-GCMS) were used to study changes in lignin which occur on down regulation of two genes which form part of the biosynthetic route to S and G lignin. Antisense down regulated p-coumourate 3-hydroxylase (C3H) and hydroxycinnamoyl transferase (HCT) mutants of Alfalfa (*Medicago sativa*) were used to investigate the impact of changes in total lignin content, as well as H, G and S lignin ratios. Both genes are involved in the S and G but not H biosynthetic pathways, down regulation, therefore, results in reduced production of S and G and an increased proportion of H lignin. The change in lignin content is assessed by measuring the mass spectrum of a pyrolyzed sample. Distinctive peaks are known to come from certain compounds in biomass allowing changes in those peaks to be attributed to changes in the compounds.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Integrated High Throughput Pretreatment and Enzymatic Hydrolysis in 96 Well Plates

Michael H. Studer[1]* (michael.studer@ucr.edu), Jaclyn DeMartini,[1] Heather McKenzie,[1] Charles E. Wyman,[1] and **Martin Keller[2]**

[1]Dept. of Chemical and Environmental Engineering, Center for Environmental Research and Technology, University of California, Riverside, Calif.; and [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Knowledge gained by thoroughly characterizing biomass chemistry and structure will drive coordinated development of improved plant biomass and degradation microbes. Native, genetically modified, and partially deconstructed lignocellulosic samples will be analyzed.

One of the aims of the BioEnergy Science Center (BESC) is the engineering of advanced plants with reduced recalcitrance for sugar release. As this requires screening of many different natural and genetically modified biomass types to identify those with a lower recalcitrance for sugar release, a new high-throughput (HTP) tool integrating pretreatment and enzymatic hydrolysis in the same multi-well configuration was developed. Water only or dilute acid pretreatment of 1% biomass slurries is performed in a custom-made 300µL 96 well plate made of metal to withstand heating to temperatures up to 180˚C in a steam chamber as well as prevent corrosion. Furthermore, our so-called co-hydrolyis approach adds citric acid buffer, sodium azide, and enzymes (e.g., cellulase and xylanase) directly to each well without separating the solid and liquid after pretreatment, with enzyme loadings based on the original glucan and xylan content of the raw biomass. Next, the plate is incubated at 50˚C for 72 h, and the release of sugars is quantified by HPLC. To prove the feasibility of this concept, performance for the co-hydrolysis process was compared to that for conventional washed solids hydrolysis, and about 100 mg

of enzyme protein/g glucan and xylan in the raw biomass was shown to overcome inhibition by products formed or released during pretreatment and result in similar yields. Operational testing demonstrated that the custom-made well-plates did not leak during pretreatment, and heat-up and cool-down required less than 45 s at an operating temperature of 180°C. The standard deviation in total sugar yields was only 4.1% across the 96 wells for combined pretreatment and co-hydrolysis of poplar, and yields for co-hydrolysis using the multi-well system were virtually identical to those for co-hydrolysis with standard tube reactors as well as washed solids hydrolysis using these standard vessels.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Advanced Imaging Projects in the BioEnergy Science Center (BESC)

Shi-You Ding[1]* (shi_you_ding@nrel.gov), Yining Zeng,[1] Thomas Haas,[1] Yu-San Liu,[1] Mike Himmel,[1] X. Sunney Xie[2] (xie@chemistry.harvard.edu), and **Martin Keller**[3]

[1]Chemical and Biosciences Center, National Renewable Energy Laboratory, Golden, Colo.; [2]Dept. of Chemistry and Chemical Biology, Harvard University, Cambridge, Mass.; and [3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Knowledge gained by thoroughly characterizing biomass chemistry and structure will drive coordinated development of improved plant biomass and degradation microbes. Native, genetically modified, and partially deconstructed lignocellulosic samples will be analyzed.

Deeper understanding of biomass structure and processes for its conversion to sugars and other useful products is required to enable economic lignocellulosic biofuels production. The BioEnergy Science Center (BESC) has targeted these objectives for critical investigation. Imaging techniques are used to study the cell walls from wild type, genetically modified, and chemically and/or biologically treated switchgrass and poplar to identify the molecular characteristics that govern the emergent property of recalcitrance. Our projects focus on four research areas:

Sub-Nanometer Scale Imaging of the Plant Cell Walls: The structural networks in higher plant cell walls are composed of polysaccharides, pectins, lignins and glycoproteins, which are evolved to be recalcitrant composites that are difficult to deconstruct to fermentable sugars. It remains a challenge; however, to characterize plant cell walls at the molecular scale. New imaging tools based on scanning probe microscopy (SPM) are capable of characterize the plant cell wall at the nanometer scales. Specifically, atomic force microscopy (AFM) combined with near-field optics is now used to image cell walls with considerable success.

High Resolution Chemical Imaging of Biomass: It is important to identify the chemistry and kinetics of the biomass conversion processes *in vivo*. Non-linear optical microscopy has been used to map plant cell wall chemistry in the BESC. Specifically, coherent anti-Stokes Raman scattering (CARS), stimulated Raman scattering (SRS) (see also Poster by *Saar et al.*), and second (SHG) and third (THG) harmonic generation are currently used to enable new imaging tools for the characterization of biomass samples.

Microbe/Enzyme/Biomass Interaction: In nature, biomass degradation is a process of molecular interaction and reaction between plant cell wall polymers (i.e., cellulose and matrix polymers) and cellulolytic microbes and their secreted enzymes. An integrated system has been set up to combine microscopic and spectroscopic modules that allow us to characterize biomass conversion processes at high spatial and chemical resolution. For example, AFM is used to map the surface topography of the plant cell wall and the binding of microbial cells and enzymes to the walls; total internal reflection fluorescence (TIRF) microscopy with fluorescent probe labeling techniques is used to track the distributions and movements of labeled microbial cells and enzymes; and spectroscopy is used to monitor the resultant chemical changes in cell wall polymeric component during biochemical, as well as chemical conversions of biomass.

Single Enzyme Molecule Spectroscopy: It is believed that efficient deconstruction of complex plant cell walls requires the synergistic reaction of many polysaccharide-degrading enzymes. In some cases multiple enzyme complexes (i.e., cellulosomes) perform the efficient hydrolysis of the plant cell walls. It is not well known; however, the ways that these enzyme complexes are assembled and made functional. We have shown that single molecule approaches are capable of tracking the behaviors of individual enzyme and binding domains known to be critical for the cellulase function. At NREL, various carbohydrate-binding modules and enzymes have been tagged with fluorescent probes so that TIRF and non-linear optics can be used to image these molecules at the single molecular level.

We note that integrated imaging modalities are required to characterize biomass, both in the wild type form and during the conversion process. The knowledge derived from the use of these advanced characterization tools could be facilitated by combining our ability to image cell walls at the meso- and micro- scales, i.e., by the simultaneous imaging of single molecule features upon a complex chemical background. Example images of plant tissues, cellulose microfibrils, microbial cells, enzyme complexes, and single enzymes will be presented and discussed.

References

1. Ding S-Y and & Himmel ME, (2006) The Maize Primary Cell Wall Microfibril: A New Model Derived from Direct Visualization. *J. Agric. Food Chem.* 54, 597-606.
2. Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, and Foust TD (2007) Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. *Science* 315, 804-807.
3. Ding S-Y, Himmel ME and Xie SX, (2008) Identify Molecular Structural Features of Biomass Recalcitrance Using Nondestructive Microscopy and Spectroscopy. *Microscopy and Microanalysis*, Vol. 14, Suppl. S2, pp 1494-1495 doi: 10.1017/S1431927608088582.
4. Ding S-Y, Xu Q, Crowley M, Zeng Y, Nimlos M, Lamed R, Bayer EA and Himmel ME, (2008). A biophysical perspective on the cellulosome: new opportunities for biomass conversion *Curr. Opin. Biotechnol.* 19:218–227.
5. Freudiger CW, Min W, Saar BG, Lu S, Holtom GR, He H, Tsai JC, Kang JX, Xie SX, (2008). Label-free Biomedical Imaging with high sensitivity by stimulated Raman scattering microscopy, *Science,* in press.
6. Xu Q, Tucker MP, Arenkiel P, Ai X, Rumbles G, Sugiyama J, Himmel ME, and Ding S-Y, (2008) Labeling the planar face of crystalline cellulose using quantum dots directed by type-I carbohydrate-binding modules. *Cellulose*. DOI 10.1007/s10570-008-9234-4
7. Ding S-Y, Xu Q, Ali MK, Baker JO, Bayer EA, Barak Y, Lamed R, Sugiyama J, Rumbles G, and Himmel ME, (2006). Versatile derivatives of carbohydrate-binding modules for imaging of complex carbohydrates approaching the molecular level of resolution. *Biotechniques* 41, 435-443.

This work is funded by the US Department of Energy, the Office of Science, Biological and Environmental Research (BER) program under two projects, DOE Genomics:GTL BioEnergy Science Center (BESC) and "Study of Lignocellulosic Material Degradation with CARS Microscopy" Grant DE-FG02-07ER64500.

GTL

Analytical BESC Advances in Characterization of Biomass and Recalcitrance

Y. Pu[1]* (Yunqiao.Pu@ipst.gatech.edu), N. Jiang,[1] H. Li,[1] M. Foston,[1] R. Samuel,[1] J. Seokwon,[1] A.J. Ragauskas,[1] and **Martin Keller**[2]

[1]School of Chemistry and Biochemistry, Institute of Paper Science and Technology, Georgia Institute of Technology, Atlanta, Ga.; and [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Knowledge gained by thoroughly characterizing biomass chemistry and structure will drive coordinated development of improved plant biomass and degradation microbes. Native, genetically modified, and partially deconstructed lignocellulosic samples will be analyzed.

The design of low recalcitrance biomass for deconstruction to bioethanol and related biofuels is predicated on understanding the fundamental relationship between plant polysaccharides, lignin and how these biopolymers are integrated in the plant cell wall. This poster examines how advanced 1D and 2D NMR techniques can be utilized to identify structural elements of importance to the recalcitrance of genetically engineered alfalfa. In addition, on going studies utilizing deuteronic ionic liquid solvents with NMR and MALDI techniques are being developed such that rapid whole plant cell wall structures can be determined without the need for laborious isolation and purification techniques. We will also report crystallinity and other measurements as part of understanding the structure and chemistry of these samples. These advances will provide researchers new analytical tools to address the challenge of high-throughput/high resolution analysis of biomass polymers generated in state-of-the-art plant genomics programs.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Computer Simulation of Lignocellulosic Biomass

Loukas Petridis[1]* (petridisl@ornl.gov), Benjamin Lindner,[1] Jeremy C. Smith,[1] and **Martin Keller**[2]

[1]Center for Molecular Biophysics and [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Combining experimental data with biological models will help define the genomic and physical basis of plant cell-wall recalcitrance and deconstruction.

Plant cell wall lignocellulosic biomass is a complex material composed of crystalline cellulose microfibrils laminated with hemicellulose, pectin, and lignin polymers. The aim of this project is to use computer simulation to complement experiments in understanding the physical properties of this biomass. Atomistic models of softwood lignin and cellulose have been created and combined to form an initial model of lignocellulose. The models were built based on input from experiments on lignin and cellulose composition from the characterization group of BioEnergy Science Center. Molecular dynamics simulation of the models is performed to examine for structural properties of the system.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

The Plant Cell Wall Biosynthesis Related Galacturonosyltransferase (GAUT) and GAUT-Like (GATL) Genes Have a Different Origin than the Other Glycosyltransferase Family 8 Genes

Yanbin Yin[1,3]* (yinyb@csbl.bmb.uga.edu), Huiling Chen,[1] Michael Hahn,[2,3] Debra Mohnen,[2,3,] Ying Xu,[1,3] and **Martin Keller**[3]

[1]Computational System Biology Laboratory, Dept. of Biochemistry and Molecular Biology, and Institute of Bioinformatics, and [2]Complex Carbohydrate Research Center, University of Georgia, Athens, Ga.; and [3]BioEnergy Science Center, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Combining experimental data with biological models will help define the genomic and physical basis of plant cell-wall recalcitrance and deconstruction.

The galacturonosyltransferases (GAUT) and GAUT like (GATL) genes, belonging to the glycosyltransferase family 8 (GT8), have been indicated to be involved in the plant cell wall biosynthesis (pectin and xylan synthesis) (see poster by Z. Hao et al.). In this study, we have identified, classified and evolutionarily studied the GT8 family in 13 fully sequenced plant and green algal genomes as well as in the NCBI non redundant protein database (NCBI-nr DB). We found that there are three major GT8 protein classes in nature. The GAUTs and GATLs are in class I, while galactinol synthases (GolSs) and plant glycogenin-like starch initiation proteins (PGSIPs) are in class II. The class III includes almost all bacterial sequences. A moss-specific subfamily, a metazoan subfamily, some algal sequences and a sequence from the cyanobacteria *Synechococcus elongatus* PCC 7942 are found within class I too. Some sequence motifs such as the DXD and HXXGXXKPW are conserved across all GT8

* Presenting author

subfamilies, while others are specific to certain subfamilies. We conclude that the plant cell wall synthesis related GAUTs and GATLs have a different origin than the other plant GT8 genes, possibly anciently acquired from some certain cyanobacteria.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

Computational Prediction of Golgi Resident Proteins in *Arabidopsis thaliana*

Wen-Chi Chou[1,3]* (wcc@uga.edu), Maor Bar-Peled,[2,3] Ying Xu,[1,3] and **Martin Keller**[3]

[1]Computational Systems Biology Laboratory, Dept. of Biochemistry and Molecular Biology, and Institute of Bioinformatics and [2]Center for Complex Carbohydrate Research and Dept. of Plant Biology, University of Georgia, Athens, Ga.; and [3]DOE BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Combining experimental data with biological models will help define the genomic and physical basis of plant cell-wall recalcitrance and deconstruction.

Proteins residing in plant Golgi play key functional roles in cell wall synthesis and protein targeting. It will provide a much improved understanding about the detailed mechanism of cell wall syntheses if we know which functional proteins reside in Golgi. We present here a computational method for identification of the Golgi resident proteins in *Arabidopsis thaliana*. We have compiled a list of Golgi residing proteins from the published literature, and studied a number of sequence and structural features of these proteins, potentially useful in distinguishing between Golgi resident and non-Golgi proteins. Among these features, we found that transmembrane domains and a few functional signatures associated with known Golgi resident proteins are particularly useful. Based on these features, we have developed a classification method based on a support vector machine, for predicting Golgi resident proteins. Our prediction program achieved 91% and 93% prediction sensitivity and specificity, respectively, on a test set consisting of 1,434 proteins of *Arabidopsis thaliana*, substantially better than any of the existing prediction programs. Using this program, we have predicted around 2700 Golgi resident proteins in *Arabidopsis thaliana*, providing a useful data source to other researchers.

The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

The BioEnergy Science Center Laboratory Information Management System (LIMS)

Susan K. Holladay* (holladaysk@ornl.gov), Denise D. Schmoyer, Sheryl A. Martin, Edward C. Uberbacher, and **Martin Keller**

Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Combining experimental data with biological models will help define the genomic and physical basis of plant cell-wall recalcitrance and deconstruction. Laboratory Information Management Systems and knowledgebase environments have separate key roles. The LIMS is the central data repository for all information related to processing and analysis of the BioEnergy Science Center (BESC) samples. It maintains a detailed history for each sample by capturing processing parameters, protocols, statistical data, stocks, tests and analytical results for the complete life cycle of the sample. It is also used to track protocols so that data can be linked to its source and history. The LIMS is also linked to the BESC knowledgebase which

contains the results of experiments interpreted in the context of biological systems.

A laboratory information management system has been implemented using the Nautilus commercial package and modified and configured for BESC operations. The major data generation steps used by BESC experimental campaigns in Year 1 have been coded into this LIMS system. These include data and metadata from various instruments and processes covering sample acquisition, biomass grinding, biomass pretreatment, compositional analysis before and after cellulolytic processing, the analysis of complex environmental community microbial samples, and sequence analysis using 454. The LIMS has been used to capture results and process for the eight campaigns undertaken by BESC in Year 1. For each of these campaigns a detailed workflow, including sample generation, sample splitting and shipping, sample processing, protocols, controls, replicates, and metadata and results have been captured. This has involved extensive collaboration and communication between the LIMS team and experimental labs. The tracking of materials, samples and processes is central to the proper functioning of BESC both from the standpoint of ensuring meaningful and reproducible results, but also to track the generation of intellectual property. All materials and samples shipped between institutions are tracked at the level of the institution using a material transfer information system attached to the LIMS. The LIMS has been used to capture large amounts of primary data including over 800 results from compositional analysis studies. The LIMS team has provided initial documentation on LIMS use for sample tracking on the Wiki site. Additionally several seminars about LIMS use have been provided with additional training workshops to be held in Year 2. Other outreach mechanisms including mailings and helpline have been provided. An ORACLE database has been designed and implemented to support LIMS operations. It stores data, metadata and links to data files from to BESC experiments and procedures. The database server has been set up in a special area of the ORNL network outside the firewall (Open Research Enclave), which also houses the LIMS system and interfaces. Processes have been established for user account and password generation that ensure proper data security. Using the database and available LIMS interface, users can access data about samples or analysis results from any campaign.

A standardized approach to bar coding has been implemented in BESC, with bar codes generated by the LIMS system. Bar codes are used to initiate (generate) samples and also when samples are sub-sampled, or used in multi-well plate operations as specified in each workflow. In addition to unique IDs contained in the LIMS, human readable fields are available on the barcodes. Major labs in BESC have all purchased the same barcode reader purchased as a result of substantial testing by NREL. A series of reports is now available for BESC members including reports for material transfer, shipping, sample history and provenance, campaign results.

This submission has been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

The BESC Knowledgebase: An Infrastructure for Biological Discovery

Morey Parang[1,2*] (parangm@ornl.gov), Tatiana Karpinets,[1,2] Doug Hyatt,[1,2] Mike Leuze,[1] Honney Park,[1] Bob Cottingham,[1] Michael Galloway,[1] Loren Hauser,[1] Mustafa Syed,[1] Ed Uberbacher,[1] and **Martin Keller**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [2]University of Tennessee, Knoxville, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multi-talented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). BESC researchers in characterization, modeling, education and data management seek to provide integrating data and understanding across the center. We use advanced technologies to analyze chemical and structural changes within biomass. We have developed a coordinated characterization pipeline to provide composition, pretreatment, and enzyme digestibility data from native and modified plant samples and to store this data within a data management system. We are developing knowledge environments to allow access to improved data analysis and modeling tools. Combining experimental data with biological models will help define the genomic and physical basis of plant cell-wall recalcitrance and deconstruction. Laboratory Information Management Systems and knowledgebase environments have separate key roles.

The BioEnergy Science Center (BESC) is a multi-institutional center which is undertaking large experimental campaigns to understand and reduce biomass recalcitrance. As such, the center is expected to generate large volumes of diverse data including genome sequences, omics data, protein structures, images, mass spectrometry as well as NMR spectra, and various assay results. The BESC knowledgebase (KB), currently under development, provides a novel means for creative exploration and sharing of the data, materials, and experimental processes across the distributed enterprise. The BESC KB is an infrastructure designed specifically for providing "knowledge" of BESC-related research. It serves as a participatory biological discovery platform for the BESC community. The BESC KB links and correlates the results of analyses from BESC analysis processes to

* Presenting author

genomes, pathways, regulatory knowledge and molecules from plants and microbes. The KB platform provides an array of applications and tools to extract and combine knowledge from isolated data, helping the BESC researchers in interpretation and understand of the experiential results and to form rapid hypotheses (Figure A).

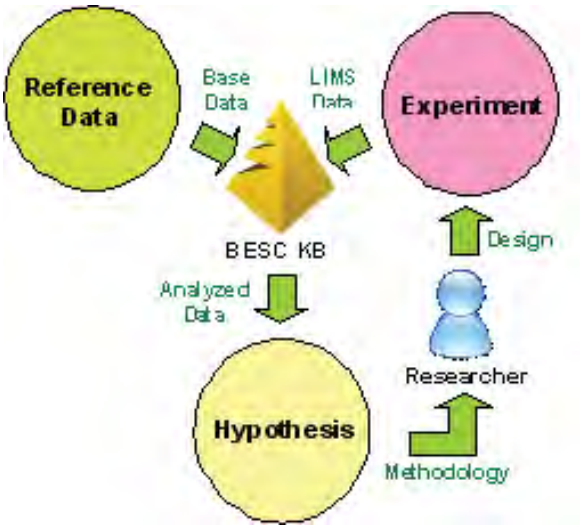


Figure A. BESC Knowledgebase Concept and Rational

Figure B. BESC Knowledgebase BeoCyc

An ultimate goal of the BESC knowledgebase is to develop an infrastructure that (i) provides comprehensive information on metabolic processes in the organisms related to production of biofuel, (ii) facilitates analysis and querying of the information in a cell systems context and (iii) overlays the information with the collected experimental data and other information in the BESC knowledgebase.

The knowledgebase addresses the main BESC focus areas through development of five "knowledge" domains: (i) Plant Genomes and Cell Wall Biosyntheses; (ii) Cellulolytic Microbes; (iii) Biomass Characterization Data; (iv) Enzymes and Cellulosomes; (v) Consolidated Bioprocessing. Currently only the first two domain areas have been implemented. The Plant Genomes and Cell Wall Biosyntheses component primarily targets the Biomass Formation and Modification focus area knowledge requirements. This domain contains reference data for twenty one plant genomes along with a diverse set of bioinformatic tools and applications. These tools include annotation pipelines, manual curation of gene models, visualization tools, and search engines.

The Cellulolytic Microbes domain targets the Characterization and Modeling focus area, which includes the BeoCyc implementation. BeoCyc is an initial element of the infrastructure (Figure B) representing a collection of Pathway/Genome Databases (PGDBs) in a set BioEnergy related microOrganisms (Beo). The PGDBs were generated automatically using the Pathway Tools software and include annotation of genes in the microorganisms with metabolic pathways from the MetaCyc database. BeoCyc provides a web interface to the PGDBs and to diverse computational, visualization and analytical tools based on the generated databases, which include searching of the databases, browsing of the microbial genomes and generated annotations, downloading the annotations, overview of the metabolic pathways and their overlaying with the experimental data. At present BeoCyc includes PGDB reconstructions for 14 microbial organisms involved in degradation of cellulose and fuel productions and for some bacteria that are specific targets of BESC investigators.

The submitted abstract has been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

GTL

BioEnergy Science Center Education and Outreach

P. Doney[1,3]* (patdoney@uga.edu), S. Fowler,[2,3] B. Davison,[2,3] **M. Keller,**[2,3] J. Westpheling,[1,3] S. Kral,[4] J. Griffin,[4] and H. Schulson[4]

[1]University of Georgia, Athens, Ga.; [2]Oak Ridge National Laboratory and [3]BESC, Oak Ridge, Tenn.; and [4]The Creative Discovery Museum, Chattanooga, Tenn.

http://bioenergycenter.org

Project Goals: The BioEnergy Science Center (BESC) is focused on the fundamental understanding and elimination of biomass recalcitrance. BESC's approach to improve accessibility to the sugars within biomass involves 1) designing plant cell walls for rapid deconstruction and 2) developing multitalented microbes for converting plant biomass into biofuels in a single step (consolidated bioprocessing). Education and outreach to

the general public is critical in the acceptance and deployment of bioenergy. In addition to leveraging successful education and training programs already in place at our partner institutions, BESC has developed educational lessons and activities that target elementary and middle school children.

In addition to our efforts to prepare a new generation of scientists for the emerging fields of bioenergy through the interdisciplinary training of graduate students and postdocs, our center has taken a novel approach in that our education efforts begin with fifth graders. We have developed lesson plans aimed at 4th, 5th and 6th grades to educate and inform students about the basics of energy production and utilization. They include basic concepts such as the carbon cycle, lignocellulosic biomass as substrate for the production of biofuels as well as technical and economic obstacles to a biobased fuel economy. The hands-on activities and guided questions are also designed to meet educational objectives for these grades. These lessons have been piloted in a hundred classrooms in North Georgia and Tennessee and will be made available to schools nationwide in the fall of 2009. We have also began to pilot interactive "science night" programs offered to students and the general public through local schools, museums and community centers. We will present details of some of the lessons and science night activities.

The BioEnergy Science Center is supported by the Office of Biological and Environmental Research in the DOE Office of Science.

USDA–DOE Plant Feedstock Genomics for Bioenergy

USDA-DOE

Towards a Map of the *Populus* Biomass Protein-Protein Interaction Network

Eric Beers[1]* (ebeers@vt.edu) Amy Brunner,[2] Allan Dickerman,[3] Xiaoyan Sheng,[2] Chengsong Zhao,[1] and Mingzhe Zhao[1]

[1]Dept. of Horticulture, [2]Dept. of Forestry, and [3]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Va.

http://xylome.vbi.vt.edu/

Project Goals: 1) Clone approximately 400 Gateway-compatible ORFs corresponding to the poplar xylem gene set (PxORF); 2) Identify poplar proteins that co-purify with selected TAP-tagged PxORF proteins expressed in poplar; 3) Identify high-confidence Y2H interactions for a subset of approximately 60 PxORFs comprising putative regulators of lignocellulose synthesis screened against a poplar xylem cDNA library; 4) Identify Y2H interactions resulting from a matrix of pair-wise assays between all PxORF proteins; 5) Produce a protein-protein interaction map that incorporates interactions identified from the three screens; and 6) Maintain a web site to make results available and facilitate distribution of clones.

As one of the fastest-growing and most productive trees in North America, poplar is a model biomass crop for producing "green" energy such as power, heat, and biofuels. The lignocellulosic (woody) portion of the plant is a sink for CO_2, and can be converted to ethanol to provide a renewable energy source. Therefore, a more detailed understanding of the molecular biology and genomics of wood formation in poplar trees is needed for the development of novel strategies, which will ultimately lead to the creation of low-cost and more efficient bioenergy feedstocks. The poplar genome is the first completely sequenced tree genome containing over 45,000 protein-coding genes. Our previous microarray data have shown that approximately 400 poplar genes are highly up-regulated in xylem. These genes have been designated as the poplar biomass gene set. In this project, our overall goal is to identify protein-protein interaction networks associated with biomass production in the woody tissues of poplar by performing yeast two-hybrid matrix assays using all possible pair-wise combinations from the biomass gene set tested in the same yeast cell for the strength of their interaction. Furthermore, selected genes will be tested in the yeast two-hybrid system for their abilities to interact with any protein found in a library of proteins derived from all the genes expressed in poplar wood-forming tissues. Finally, selected genes will be expressed in plants or *E. coli* as affinity-tagged proteins and tested for their abilities to form complexes with other proteins extracted from woody tissue. Members of protein complexes isolated from plant extracts will be identified using mass spectrometry. All of the interactions revealed as a result of these three types of protein-protein interaction studies will be combined and processed by bioinformatical programs to produce a biomass protein-protein interaction map. Significant progress has been made in constructing our poplar biomass ORF collection. To date, we have successfully cloned into Gateway entry vectors more than 120 open reading frames (ORFs) that encode members of the poplar biomass gene set. Sequencing of these biomass ORFs has revealed numerous splice variants and allelic sequences that diverge from the JGI model ORFs. Approximately 90 ORFs have been recombinationally cloned into yeast vectors as DNA-binding domain (DB) and activation domain (AD) fusions for two-hybrid assays. All DB fusions are tested for auto-activation of reporter genes without the presence of an AD fusion vector. Auto-activating DB fusions are eliminated from future yeast two-hybrid assays or re-cloned minus their putative activation domains. A pilot 30 x 30 yeast two-hybrid matrix experiment is being performed focusing on genes involved in secondary cell wall synthesis and signal transduction. Eleven genes have been affinity-tagged and transformed into poplar. Independent transgenic poplar lines have been tested by real-time RT-PCR and/or Western blotting to confirm the expression of affinity-tagged proteins.

USDA-DOE

Development of Genomic and Genetic Tools for Foxtail Millet, and Use of These Tools in the Improvement of Biomass Production for Bioenergy Crops

Jeffrey L. Bennetzen[1]* (maize@uga.edu), Katrien M. Devos[2] (kdevos@uga.edu), Andrew Doust[3] (andrew.doust@okstate.edu), Janice Zale[4] (janicezale@comcast.net), and Ryan Percifield[1]

[1]Dept. of Genetics, [2]Dept. of Crop and Soil Science, and Dept. of Plant Biology, University of Georgia, Athens, Ga.; [3]Botany Dept., Oklahoma State University, Stillwater, Okla.; and [4]Dept. of Plant Sciences, University of Tennessee, Knoxville, Tenn.

Project Goals: This project will generate a variety of genomic and genetic tools for foxtail millet, including SNPs, BAC libraries, optimized foxtail millet transformation technology, and a high density QTL and genetic map of foxtail millet for significant biomass traits. These resources will complement the DOE Joint Genome Institute whole genome sequencing of foxtail millet, enhancing its value as a functional genomic model for second generation bioenergy crops such as switchgrass.

Foxtail millet, *Setaria italica*, has the full set of attributes that will make it a model plant for basic and applied studies, particularly for its close relatives like switchgrass, an important bioenergy crop. Foxtail millet has small and simple chromosomes that will be sequenced in 2009 by the DOE's Joint Genome Institute (JGI). This sequence will enormously enhance further study and understanding of the genetic basis of biomass production, but it will only be fully useful if additional technologies are developed for use in foxtail millet. This proposal described experiments to improve the genetic and genomic toolkit for foxtail millet, and thus assist with the assembly of the foxtail millet genomic sequence, anchor the sequence to the genetic map and lay the foundation for future molecular, developmental, physiological and genetic studies.

The objectives and approaches in this proposal were 1) to construct a >12X redundant clone (BAC) library from foxtail millet inbred Yugu1 DNA, the DNA that will be sequenced by JGI, 2) to sequence the ends of the BACs so these data can help in the assembly and annotation of the complete foxtail millet genome sequence, 3) to confirm the assembly and detailed annotation, 4) to generate a new and large mapping population from the cross of Yugu1 foxtail millet with its wild progenitor, *Setaria viridis*, 5) to sequence the *S. viridis* genome with next-generation technology to find single nucleotide polymorphism (SNP) markers, 6) to construct a >1000 marker genetic map from the new crossing population using these SNPs, 7) mapping of tillering, shade response and biomass yield QTL in this new population, and 8) optimization of foxtail millet transformation technology and its initial use to generate transgenics with altered expression of biomass candidate genes.

As of December 1 of this year, three months after grant activation, BAC library construction (obj. 1) has been completed and the BACs have been sent to JGI for end sequencing (obj. 2). The assembly and annotation of the complete foxtail millet genome sequence (obj. 3) will be in collaboration with the JGI, emphasizing the high throughput expertise of the JGI group and the manual confirmation skills of the UGA group. The generation of a new and large mapping population of a cross between foxtail millet and green millet (obj. 4) has begun, with an F_2 population now in hand. Further crosses are planned to increase the F_2 population and to try other green millet accessions as parents. To help integrate the first genetic map into the planned new map, 110 previous RFLP markers are being converted into PCR-based markers.

USDA-DOE

Identifying Genes Controlling Feruloylation in Grass Cell Walls

M.M. de O. Buanafina* (mmb26@psu.edu) and D. Braun (dda10@psu.edu)

Biology Dept., Pennsylvania State University, University Park, Pa.

Project Goals: The focus of this proposal is to identify and characterize new genes controlling feruloylation in grasses, as well as new genes that are responsible for the assembly of lignin into the cell wall and for biomass conversion. This will provide fundamental knowledge concerning the most crucial factors that influence grass cell wall degradability.

The major wall phenolics in grasses include lignin and hydroxycinnamic acids (HCA). HCAs constitute 1 to 2% of grass cell walls by dry weight and are important structural components as they are ester and ether linked to arabinoxylans. Ferulic acid (4-hydroxy-3-methoxy-cinnamic acid) is the major HCA identified in both primary and secondary cell walls of grasses. Ferulic acid residues attached to arabinoxylans have the ability to form ferulate dimers functioning in cell wall cross-linking. They are also proposed to act as nucleation sites for the formation of lignin and for the linkage of lignin to the xylan/cellulose network. Such coupling reactions, which occur predominantly in grasses, significantly decrease cell wall degradability and thus work as a barrier against efficient utilization of cell walls as a source of biomass for bioenergy production.

We have shown previously that the expression of ferulic acid esterase FAE in different grass species resulted in a substantial reduction in cell-wall-esterified ferulates and diferulates. FAE was shown to impact cell wall hydrolysis, resulting in increased yield of reducing sugars by cellulase treatment as

well as increased digestibility, reinforcing the importance of feruloylation and cross linking for cell wall degradability.

Controlling the level of total feruloylation should have a direct impact on the level of cross-linking and thereby on cell wall degradation. Currently, the genes underlying arabinoxylans feruloylation have not been identified and the isolation of such genes could be of great importance in manipulating ferulates accretion to the wall in grass species. Accordingly, we have targeted the feruloylation pathway as a means of increasing the efficiency of energy production from plant biomass. To identify these genes we are taking a forward genetic approach combined with a spectroscopic screen followed by detailed genetic and phenotype analyses. We have chosen *Brachypodium distachyon* as our model grass system because of its unique characteristics such as small genome size. *Brachypodium* seeds have been mutagenized with ethyl-methane sulfonate. We propose to identify *Brachypodium* mutants of interest by high-throughput spectroscopy-based leaf assay screening for change in the level of cell-wall esterified ferulic acid. Mutation of the feruloyl transferase gene(s) should lead to accumulation of soluble ferulates – as they will not be secreted to the cell wall – and less ferulate cross-linked to xylan.

If successful, the identification of these genes in *Brachypodium* can work as a handle for gene discovery in other important grass crop species because of conservation of genome organization and gene order in grasses.

USDA-DOE

Developing Association Mapping in Polyploidy Perennial Biofuel Grasses: Establishing the Baseline on Genome-Size Variation

Denise E. Costich,[1]* Bernd Friebe,[2] Michelle E. Denton,[3] Elhan Ersoz,[3] Michael D. Casler,[4] Jerome H. Cherney,[5] and **Edward S. Buckler**[1] (dc58@cornell.edu)

[1]USDA-ARS and Institute for Genomic Diversity, Cornell University, Ithaca, N.Y.; [2]Wheat Genetics and Genomic Resources Center, Kansas State University, Manhattan, Kans.; [3]Cornell University, Ithaca, N.Y.; [4]USDA-ARS, U.S. Dairy Forage Research Center, Madison, Wis.; and [5]Dept. of Crop & Soil Sciences, Cornell University, Ithaca, N.Y.

Project Goals: 1) Assemble association panels of diverse populations and linkage populations for switchgrass and reed canarygrass (~1000 clones each); 2) Trait evaluation for key biofeedstock characteristics in these panels; 3) Develop high density SNP markers in switchgrass; 4) Genotype association panels and linkage populations in switchgrass; 5) Evaluate population structure and germplasm diversity in switchgrass; and 6) Establish association mapping and estimate marker based breeding values in switchgrass.

Perennial grasses are considered to be primary candidates for biofuel production. However, many aspects of the basic biology of these species, including the number of sets of chromosomes in their genomes (ploidy), are still poorly understood. In fact, nineteen of the twenty-two grass species that were screened by the Oak Ridge National Laboratory's Herbaceous Energy Crops Program in the 1980s are polyploid, a phenomenon where genetically related individuals possess chromosome numbers that are multiples of each other. Furthermore, ten of these species have variable ploidy, where 2-6 different levels (4X-12X) have been observed. Two of these species, Switchgrass (*Panicum virgatum*) and Reed canarygrass (*Phalaris arundinacea*), are the study organisms for our joint linkage and association mapping project. In addition to ploidy-level variation, these species also show evidence of loss and gain of individual chromosomes (aneuploidy).

This variation in genome size will have major implications for genetic marker development and mapping. For example, chromosome-number variation, resulting in gene-copy number variation, could underlie a significant proportion of the phenotypic variation in the biofuel-related traits we will be measuring. We are determining genome size in two ways, via flow cytometry to measure nuclear DNA content and phase contrast microscopy to count chromosomes, in a representative sample of switchgrass plants that will become part of our permanent live germplasm collection. Establishing the correlation between these two measurements will allow us to use the faster method (flow cytometry) to phenotype thousands of plants in our field populations, reserving the more labor-intensive method (microscopy) for outliers. We present here our progress to date in this study and provide recommendations to the scientific community developing genetic resources for biofuel grasses.

USDA-DOE

Computational Resources for Biofuel Feedstock Species

Kevin L. Childs* (kchilds@plantbiology.msu.edu) and **C. Robin Buell** (buell@msu.edu)

Michigan State University, East Lansing, Mich.

http://bfgr.plantbiology.msu.edu/

Project Goals: To provide computational tools and resources for data-mining genomic based information available for biofuel feedstock species. Through the Biofuel Feedstock Genomics Resource (http://bfgr.plantbiology.msu.edu/), we will create a comprehensive, uniform, well annotated resource for data-mining genomic data for biofuel feedstock species.

While current production of ethanol as a biofuel relies on starch and sugar inputs, it is anticipated that sustainable production of ethanol for biofuel use will utilize lignocellulosic feedstocks. Candidate plant species to be used for lignocel-

lulosic ethanol production include a large number of species within the Poaceae, Pinaceae, and Salicaceae families. For these biofuel feedstock species, there are variable amounts of genome sequence resources available, ranging from complete genome sequences (e.g. sorghum, poplar) to transcriptome data sets in the form of Expressed Sequence Tags (e.g. switchgrass, pine). While obtaining genome or transcript sequence is the initial step in a genomics-based approach to biological questions, the more challenging step in genomics the process of understanding gene function and how genes and their products confer the underlying processes/traits in plant biology. This is attributable to a large extent on two issues: a large percentage of genes within genomes have no known function and experimental approaches to determining gene function on a per gene basis are fiscally prohibitive. One method to improve our understanding of gene function is through comparative approaches in which sequence similarity is used to cross-annotate orthologs and paralogs thereby leveraging all available functional annotation data to improve the annotation of genes in species with limited annotation data. We will provide computational tools and resources for data-mining genomic based information available for biofuel feedstock species. Through the Biofuel Feedstock Genomics Resource (http://bfgr.plantbiology.msu.edu/), we will create a comprehensive, uniform, well annotated resource for data-mining genomic data for biofuel feedstock species. Functional and comparative genomics will be used to improve the quality and quantity of annotation available for biofuel feedstock species. Model genomes will be included in the comparative analyses to leverage the wealth of information and resources currently available. We will generate new hypotheses regarding gene function using a systems biology approach in which gene associations in rice are linked to orthologous sequences in other Poaceae biofuel feedstock species. Collectively, this project complements existing genomic and bioinformatic resources in that it is focused on integrating sequences and annotation data already available and leverages existing data to allow for translation of information from well understood genes and genomes to biofuel feedstock species, especially those in which only partial genome sequence, annotation and functional genomics resources are available.

USDA-DOE

Translational Genomics for the Improvement of Switchgrass

Bryan W. Penning, Anna T. Olek, Maureen C. McCann, and **Nicholas C. Carpita*** (carpita@purdue.edu)

Purdue University, West Lafayette, Ind.

Switchgrass is targeted to become a future biomass crop, but the discovery of genes underlying biomass-relevant traits is compromised in switchgrass by the paucity of genetic resources. Maize provides a genetic resource for improvement of distinct cell walls of switchgrass and other energy grasses. Comparative genomics of maize, rice, and *Arabidopsis* sequences reveal marked differences in gene family structure between grass species and dicotyledonous species. A description of function for an estimated 1500 genes related to cell wall biology will be provided for switchgrass, based on homology to maize and rice sequences, and augmented with gene families that are currently of unknown function. Maize cell wall gene family members that are highly expressed during primary and secondary wall formation will be determined, and regulatory small RNA sequences will be identified. As proof of concept, we will test mutants and transgenic lines of maize for release of glucose and xylose in functional analyses of genes that impact structure and degradability of non-cellulosic polysaccharides. Analyses of cell wall structures and saccharification potential will be applied to developing stems and stover of maize and switchgrass.

USDA-DOE

Genomic Knowledgebase for Facilitating the Use of Woody Biomass for Fuel Ethanol Production

Vincent Chiang (Vincent_Chiang@ncsu.edu), Ron Sederoff, Ho-min Chang, Ying-Hsuan Sun,* Ting-Feng Yeh, Shanfa Lu, Quanzi Li, and Rui Shi

Dept. of Forestry and Environmental Resources and Dept. of Wood and Paper Science, North Carolina State University, Raleigh, N.C.

Project Goals: We propose to develop new, genome-wide understanding of the regulatory control of cellulose, xylan and lignin accumulation in wood cell walls of *Populus*. To do this, transcriptional profiling will be applied to two *Populus* systems: (1) our existing transgenic *Populus* lines having increased cellulose as a result of lignin reductiona and (2) mechanical stress induced tension wood known for its augmented cellulose and reduced xylan and lignin formation. The combination of these two systems is unique in that it may provide information about transcriptional control for preferential distribution of carbon between cellulose, xylan or lignin. Transcriptional profiling of a biological system where the process of interested is preferentially regulated is most effective for identifying the responsible genetic control. We wish to establish a comprehensive database to include all the possible genes or gene networks that may enable effective manipulation of the biosynthesis of cellulose, xylan and lignin. Our goal is to provide genomics based approaches for developing novel, high energy woody biomass for the production of ethanol. *Populus* is the only DOE target energy crop whose genome has been sequenced (http://genome.jgi-psf.org/Poptr1/Poptr1.home. html), offering a unique opportunity for achieving this goal.

Wood, consisting of cellulose, hemicelluloses and lignin, is a major renewable lignocellulosic resource that can be used as feedstock for fuel ethanol production. However, lignin and xylan hemicellulose are two major barriers to an

efficient lignocellulose conversion to ethanol. Tension wood is a specialized wood tissue formed only in angiosperm trees species, such as *Populus*, to counteract mechanical and gravitational stresses. Tension wood contains substantially increased amount of cellulose and reduced quantity of lignin and xylan. It also contains increased number of fiber cells, of which walls are enriched with more extractable lignin (high lignin S/G component ratio). These tension wood properties are highly desirable for ethanol production. The knowledge of tension wood development provides insights into regulatory networks associated with the control of the biosynthesis of cellulose, hemicelluloses and lignin, forming the basis for strategies to create more desirable wood or lignocellulosics for fuel ethanol production.

The sequenced genome of *Populus trichocarpa* allows us to take a genome-wide approach to studying regulatory mechanisms underlying the formation of tension wood. We used one-year-old greenhouse grown *P. trichocarpa* trees to study the development of tension wood by bending the tree stem to create the mechanical stress-induced tension wood. The bending was performed for 1, 3, 7, 21, 35 and 49 days. The developing xylem tissues of the bent stem as well as stem of normal trees (control) were collected for profiling gene expression using Affymetrix *P. trichocarpa* full genome microarrays. Cell wall sugar contents (for glucose, xylose, mannose rhamonose and arabinose), xylan synthase and galactan synthase activities were quantified and correlated with microarray-derived gene expression patterns.

In addition to the expected changes, such as a sharp reduction in xylan content, we also found a drastic increase of galactan quantity during tension wood formation. Galactan, a C-6 hemicellulose, is normally present as a minor hemicellulose in wood of angiosperms. Consistently, we found increased galactan and decreased xylan synthase activities in developing xylem of tension wood. Microarray experiments identified 1220 differentially expressed genes during tension wood formation based on a statistical robust mixed-model analysis and stringent Bonferroni correction for multiple tests. These differentially expressed genes include those involved in major sugar biosynthesis pathways. Based on the Pearson's correlation coefficient between the expression profiles and galactan synthase activities, we identified 6 putative galactosyl transferases (Pearson's r range from 0.83 to 0.93) that may be responsible for the increased galactan content in tension wood. Correlation analysis, however, did not identify genes that are highly correlated (Pearson's $r > 0.8$) with the decreased xylan synthase activity. The decreased xylan content found in tension wood may suggest the involvement of a post-transcriptional control of xylan biosynthesis. A proteomic approach using 2-D PAGE LC-MS/MS was also applied to complement the transcriptome results.

Parallel experiments using transgenic *P. trichocarpa* trees with reduced lignin and increased lignin S/G ratio were also carried out to identify genes that may be associated with changes in these lignin traits. Data are being analyzed and will be discussed.

USDA-DOE

Systematic Modification of Monolignol Pathway Gene Expression for Improved Lignocellulose Utilization

Richard A. Dixon* (radixon@noble.org), Fang Chen, Rui Zhou, Lisa Jackson, Gail Shadle, and Qiao Zhou

Plant Biology Division, Samuel Roberts Noble Foundation, Ardmore, Okla.

http://www.noble.org/PlantBio/Dixon/Index.htm

Project Goals: 1) To gain a detailed understanding of the biosynthetic pathways to the lignin building blocks in alfalfa, with focus on potentially independent pathways to guaiacyl and syringyl lignin; 2) To evaluate the saccharification efficiency of cell walls from alfalfa lines with various targeted modifications to lignin content and composition; 3) To develop an unbiased forward genetic screen for *Medicago truncatula* mutants with altered levels and cellular patterns of lignification, and to begin to characterize the affected genes; and 4) To provide lignin analytical support for the project "Genetic dissection of the lignocellulosic pathway of grasses" from Kansas Sate University (Bikram Gill, PI).

Alfalfa (*Medicago sativa*) has recently been promoted as a potential bioenergy crop. Currently, it is the world's major forage legume, with an average annual value of more than $8 billion in the USA alone, supported by extensive agronomic improvement programs in the private sector. Although its yields do not approach those of the currently favored bioenergy crops switchgrass, *Miscanthus* or poplar, alfalfa is a high yielding perennial with the added benefits of nitrogen fixation, well defined agronomic practices, seed industry support, and extensive translational genomics resources through studies on the closely related model species *M. truncatula*. It also has potential bioenergy uses in rotation with corn.

The compositions and structures of lignified plant cell walls have been identified as key factors limiting effective biomass to biofuel conversion in processes where the sugar components of polysaccharides are released (saccharification) prior to fermentation to ethanol. Lignin is one of the major components contributing to the recalcitrance of plant cell walls to saccharification.

Prior to the current DOE funding, we had identified the majority of the genes involved in the biosynthesis of monolignols in *Medicago,* and had generated stably transformed alfalfa lines independently down-regulated in seven different enzymatic steps of the monolignol biosynthetic pathway (L-phenylalanine ammonia-lyase, cinnamate 4-hydroxylase [C4H], hydroxycinnamoyl CoA: shikimate hydroxycinnamoyl transferase [HCT], coumaroyl shikimate 3-hydroxylase [C3H], caffeoyl CoA 3-*O*-methyltransferase [CCoAOMT], ferulate (coniferaldehyde) 5-hydroxylase [F5H], or caffeic acid 3-*O*-methyltransferase [COMT]). This resulted in lines with wide differences in both lignin

* Presenting author

content and composition in the same genetic background. During the present granting period, we have generated additional alfalfa lines down-regulated in cinnamoyl CoA reductase and cinnamyl alcohol dehydrogenase. Analysis of the above lines revealed that reduction in lignin content, more so than altering lignin composition, can result in large improvements in cell wall saccharification efficiency, and potentially obviate the need for acid-pretreatment, thus facilitating consolidated bioprocessing [1,2]. These lines have also provided baseline "improved" material for the complete repertoire of analytical procedures available within the DOE Bioenergy Sciences Center.

Although recent models of the monolignol pathway show an essentially linear pathway as far as coniferaldehyde, earlier models proposed a more complex metabolic grid. We have shown both biochemically and genetically that COMT may function at the level of 3-*O*-methylation of caffeoyl aldehyde or caffeoyl alcohol, preferred substrates for the enzyme from *Medicago*. This raises the question of the origin of caffeoyl aldehyde. We have now demonstrated that *Medicago* possesses two distinct cinnamoyl CoA reductases, one with preference for feruolyl CoA, the other with preference for coumaroyl CoA and caffeoyl CoA (the precursor of caffeoyl aldehyde). Their kinetic properties suggest that they might function to control independent pathways to the different monolignols. This is being tested through the analysis of *M. truncatula* mutant lines harboring transposon insertions in these genes. We have also identified two *Medicago* cinnamyl alcohol dehydrogenase genes encoding proteins with substrate specificities consistent with this model.

We have developed a method for the large scale screening of *M. truncatula* transposon insertion lines [3] with altered levels/patterns of lignification, by visualization of lignin/cell wall phenolic autofluorescence in stem cross sections by UV microscopy. We will describe the visible and chemical phenotypes of selected lines that have been confirmed through analysis of homozygous progeny. We currently have around 20 confirmed mutants with interesting properties, varying from an overall reduction in lignin level to ectopic lignin deposition (lacking in some cell types but not others, or present in cell types where it is not usually found). The affected genes causing these phenotypes are being pursued through analysis of genomic sequences flanking the Tnt1 insertions coupled with segregation analysis.

Finally, we have continued to provide lignin analytical support for the project "Genetic dissection of the lignocellulosic pathway of grasses". Initially, we concentrated on plant material generated through virus-induced gene silencing of wheat monolignol pathway genes, and have more recently been analyzing stable transgenic wheat lines.

References

1. Chen, F., and Dixon, R.A. (2007) Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology* 25, 759-761.
2. Jackson, L.A., Shadle, G.L., Zhou, R., Nakashima, J., Chen, F., and Dixon, R.A. (2008). Improving saccharification efficiency of alfalfa stems through modification of the terminal stages of monolignol biosynthesis. *Bioenergy Research* 1, 180-192.
3. Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., Cayrel, A., Endre, G., Zhao, P.X., Chabaud, M., Ratet, P., and Mysore, K.S. (2008) Large scale insertional mutagenesis using Tnt1 retrotransposon in the model legume *Medicago truncatula*. *Plant Journal* 54, 335-347.

USDA-DOE

Genetic Dissection of the Lignocellulosic Pathway of Wheat to Improve Biomass Quality of Grasses as a Feedstock for Biofuels

W. Li[1]* (wli@ksu.edu), C. Bi,[1] S.K. Sehgal,[1] N. Rothe,[1] H. Trick,[1] F. Chen,[2] R.A. Dixon,[2] and **B.S. Gill**[1] (bsgill@ksu.edu)

[1]Dept. of Plant Pathology, Kansas State University, Manhattan, Kans.; and [2]Dept. of Plant Biology, Noble Foundation, Ardmore, Okla.

http://www.ksu.edu/wgrc

Project goal is to identify the major determinants of lignocellulosic degradation in grasses. Our objective is to dissect the lignin pathway of grasses using wheat as a model by coupling functional genomics and metabolite profiling. Our specific objectives are: 1) to investigate the expression of candidate genes for lignin biosynthesis; 2) to silence these genes by RNAi; 3) to develop knockout mutants of these genes and; 4) to characterize the silenced tissues and knockout mutants by metabolite profiling. Combining database mining and expression pattern studies, we identified 21 candidates for the 10 families of monolignol biosynthetic genes. Expression of these 21 genes is under strong developmental regulation and they are expressed at much higher level in the stem as compared to the leaf sheath and blade. This is correlated with higher lignin content in stem as compared to leaf. We have developed RNAi constructs for these 21 genes and obtained RNAi transgenic plants for 12 genes. Expression of the target genes was silenced by up to 90%. At the same time, we have developed a mutagenesis population of 3000 single M2 plants. The metabolite profiling of the RNAi transgenic plants and screening of lignin mutants by TILLING is under way.

USDA-DOE

Development of Genomic Tools for the Improvement of Prairie Cordgrass (*Spartina pectinata)*, a Highly Productive Bioenergy Feedstock Crop

Jose L. Gonzalez* (jose.gonzalez@sdstate.edu), Arvid Boe, and Vance Owens

Dept. of Plant Sciences, South Dakota State University, Brookings, S.D.

Project Goals: Develop SSR markers and construct a linkage map for prairie cordgrass (*Spartina pectinata*) to help in the development of this species as a biomass crop.

Considerable attention has been turned towards the production of biofuels from cellulosic biomass crops. Prairie cordgrass (*Spartina pectinata*) is a North American native perennial C4 grass which yields up to 22 metric tons/ha in South Dakota plantings, compared to 11 metric tons/ha for other C4 grasses. Due to the potential of Prairie cordgrass as a cellulosic biomass crop species, a research project has been instigated to assist in this grasses domestication and utilization.

An initial step is the production of the first molecular map of prairie cordgrass. The molecular map is being produced using microsatellite markers developed from sequenced genomic libraries enriched for the microsatellite sequences; (CA)n, (GA)n, (AAG)n and (CAG)n. Initial investigations of 100 clone sequences from the four libraries identified a 60% enrichment rate. Primers were designed for the enriched sequences and tested in a sub-sampling of 16 genotypes, identifying numerous polymorphisms. Due to the positive result from the analysis of these 100 clones, a total of 3,072 clones from the (CA)n library were sequenced producing 2475 viable sequences, of which 1678 contained simple SSR regions and were used to develop primers. In total from the initial 100 sequences from the four libraries and the 1678 sequences from the (CA)n library, 551 putative loci were identified. Primers were designed for all 551 putative loci and tested for viability and polymorphism. PCR conditions were optimized to allow high through put screening of a mapping population, thus two annealing temperatures (53°C and 55°C) were utilized. Results from the primer screening identified 97 primer pairs which were polymorphic at 53°C, 134 primer pairs which were polymorphic at 55°C, 83 primer pairs amplified but were not polymorphic and 197 which didn't amplify under the high through put conditions.

The 231 polymorphic primer pairs have been used to develop a molecular map using a pseudo test cross approach using an F_1 generation of a reciprocal cross between two open-pollinated parents, collected originally from the Red River valley. The final map will be utilized for marker assisted selection in a breeding program with SDSU forage and biomass crops breeder Dr. Arvid Boe. Utilizing association mapping techniques, previously validated by our research group, quantitative traits such as yield and disease resistance will be expeditiously incorporated into the cultivated crop.

Additional research efforts include a germplasm collection throughout the Midwestern states of the USA. The genetic diversity of this core genotype collection has been evaluated with AFLPs. At the same time a disease survey is also being conducted. Additional research being undertaken to assist the domestication of Prairie cordgrass includes a characterization of the transcriptome using 454 technology, and the development of a BAC library. These efforts complement efforts in the breeding program and in crop management techniques.

USDA-DOE

Profiling the Small RNAs of *Brachypodium distachyon*

Matthias Ganssmann,* Marcelo German, Blake C. Meyers, and **Pamela J. Green** (green@dbi.udel.edu)

Delaware Biotechnology Institute, University of Delaware, Newark, Del.

Project Goals: Our project aims to examine small RNA regulation in *Brachypodium distachyon* in key tissues and in seedlings under different abiotic stresses. By profiling small RNAs and mRNAs, and investigating the targets of miRNAs, this project will enhance the value of the genome and EST information now being generated. Ultimately, this knowledge should improve our ability to devise strategies to enhance growth and yield of grasses that are less amenable to functional genomic analysis.

MicroRNAs (miRNAs) are a class of small (~21-24 nucleotide) non-coding RNAs with a major role in regulating gene expression in most multicellular eukaryotic organisms. Base pairing of a miRNA to its target mRNA prevents the translation of the mRNA either by cleavage and subsequent degradation of the transcript, or translational repression by a yet unknown mechanism. Important biological processes such as structural organization of leaves and inflorescences are regulated by miRNAs. Recent findings, mainly from the model plant *Arabidopsis thaliana*, revealed the association of miRNAs with abiotic stress responses. The goal of this study is to gain insight on small RNA mediated gene regulation in *Brachypodium distachyon* (*Brachypodium*). The close relation to major cereal grain species and to Poaceous species important for energy production, together with *Arabidopsis*-like features accentuates the suitability of *Brachypodium* to serve as a model organism for temperate grasses in functional genomic research.

Our project has four specific aims: 1. Profile small RNAs and mRNAs from key tissues, stresses, and small RNA biosynthesis mutants. 2. Characterize the regulation and target RNA cleavage products of selected miRNAs that are

associated with specific stresses. 3. Compare the *Brachypodium* data to more limited information from switchgrass to identify conserved miRNAs. 4. Provide public access to the data through a web interface.

Completion of our project objectives should result in the identification of miRNAs and target RNAs that may have important roles in stress responses in *Brachypodium* and related grasses that can be studied in mechanistic detail in the future. In addition to providing the initial results and tools for the latter, by profiling miRNAs, other small RNAs, and mRNAs very deeply, this project will greatly enhance the value of the genome and cDNA sequences now being generated. These resources can then be examined for small RNAs that have the potential to regulate most any plant process.

We have prepared small RNA libraries from major tissues of *Brachypodium* and subjected these to deep sequencing with Illumina's sequencing by synthesis method. This has yielded at least 4 million reads from each tissue and we are using this data to identify the small RNAs in the population and their abundances. The results from this initial profiling analysis will be presented and the miRNAs exhibiting the strongest differential regulation in key tissues will be identified. We have also generated mRNA libraries from several tissues and small RNA libraries from plants subjected to several environmental stresses. These are currently in various stages in our sequencing and analysis pipeline. In addition, three small RNA libraries have been sequenced from switchgrass under an NSF project that will allow conserved miRNAs to be identified.

Supported by grant DE-FG02-07-ER64450

USDA-DOE

Transcription Factor Protein Interactions with Cell Wall Gene *cis*-Regulatory Regions and Their Overall Role in Bioenergy Feedstock Properties

Naomi Young,[1] Ghislain Breton,[2] Todd C Mockler,[3] Steve A Kay,[2] and **Samuel P Hazen**[1]* (hazen@bio.umass.edu)

[1]Dept. of Biology, University of Massachusetts, Amherst, Mass.; [2]Section of Cell and Developmental Biology, University of California San Diego, La Jolla, Calif.; and [3]Center for Genome Research and Biocomputing and Dept. of Botany and Plant Pathology, Oregon State University, Corvallis, Ore.

Project Goals: The long-term goal of this project is to understand the regulatory mechanism underlying plant cell wall biosynthesis at the transcriptional level. Results from these pursuits will ultimately lead to a better understanding of growth and development and cell wall biosynthesis as well as a systems level insight into regulatory networks affecting monocot and dicot bioenergy-related properties.

There is extensive heterogeneity in cell wall composition among and within species and certainly within a single plant. Walls are modified to fulfill the particular needs of a cell type at a specific developmental stage. There should indeed be several orders of complexity to control different pathways leading to the appropriate synthesis and assembly of wall components leading to a particular wall function. We have begun to screen for transcription factor (TF) protein interactions with promoters of key cell wall genes involved in cellulose, hemicellulose, and lignin synthesis. In a pilot screen using a subset (n= 250) of the complete library of *Arabidopsis* TFs, we identified novel interactions between the Trihelix and GRAS families of TFs and the phenylpropanoid metabolic grid. Our current strategy targets the direct regulators of processive as well as the non-processive glycosyltransferases. As a follow-up of the yeast one-hybrid assay, we will use ChIP-seq to capture and identify the DNA bound to the identified transcription factor proteins using overexpressed tag versions and define the extent of their genomic targets. Concurrently, we will develop gain- and loss-of-function mutants in the dicot model *Arabidopsis* and the monocot model *Brachypodium* to determine the effects of these perturbations on cell wall properties and feedstock quality. Results from these pursuits will ultimately lead to a better understanding of growth and development and cell wall biosynthesis as well as a systems level insight into regulatory networks affecting monocot and dicot bioenergy-related properties.

USDA-DOE

Biochemical Genomics of Wood Formation: *O*-acylesterification for Alteration of Lignocellulosic Property

Chang-Jun Liu* (cliu@bnl.gov), Xiao-Hong Yu, and Jin-Ying Gou

Biology Dept., Brookhaven National Laboratory, Upton, N.Y.

Project Goals: The goals of the project are to systemically characterize plant acyl-CoA dependent acyltransferases and acylesterases; by which, to identify the specific enzymes involved in lignocellulosic acylesterification, and further to explore the underlying biochemical mechanisms and biological functions of cell wall acylesterification.

Enzymatic O-acylesterification is a common modification of plant lignocelluloses. Acylesterification affects cell wall's structural property, degradability, and eventually, the growth, development and reproduction of plant feedstocks. The enzymatic O-acylation and deacylation reactions also occur in the formation of a variety of secondary metabolites that are required for disease resistance and forming heart wood in trees. Despite the wide occurrence of O-acylesterification

in plant metabolism and cell wall modification, the enzymes involved in such structural modification remain largely elusive, and the biological roles of different types of cell wall acylesterification are unclear. We are employing biochemical genomics and reverse genetics approaches to systemically characterize plant acyl-CoA dependent acyltransferases; by which, we are going to identify the specific enzymes involved in lignocellulosic acylesterification, and further to explore the underlying biochemical mechanisms and biological functions of cell wall acylesterification.

In previous fiscal years, we have identified 94 and 61 putative acyl-CoA dependent acyltransferase and 10 acylesterase genes from poplar and *Arabidopsis,* respectively. To further probe gene expression pattern, we conducted "in silico" transcript analysis and RT-PCR for all of the identified putative poplar genes by using RNAs prepared from poplar leaf, root, developing stem, apical bud, cortex of bark, phloem, developing wood and lignified wood. The bioinformatics study allowed us to reveal unique features of acyltransferase family genes in their gene structure, organization and distribution on the genomes of both *Arabidopsis* and poplar; and the gene expression profiling resulted in the discovery of a number of acyl-CoA dependent *O*-acyltransferase and acylesterase genes preferentially expressed in secondary cell wall forming tissues. These studies laid a strong basis for understanding the evolutionary relationship among acyltransferase family genes, and the gene functions implicated in plant growth, development and metabolism.

In order to understand the molecular mechanism of cell wall acylesterification, we analyzed the wall-bound acylesters of poplar by using Liquid Chromatography-Mass Spectrometry (LC-MS), Fourier Transform-InfraRed (FT-IR) microspectroscopy, and synchrotron InfraRed (IR) imaging facility. The results revealed that the cell wall of dicotyledonous poplar, as the walls of many monocot grasses, contains a considerable amount of acylesters, primarily acetyl and *p*-hydroxycinnamoyl molecules. The "wall-bound" acetate and phenolics display a distinct tissue specific-, mechanical stress responsible- and developmental accumulation pattern, indicating distinct roles of different "wall-bound" acylesters in poplar cell wall structural construction and/or metabolism of cell wall matrix components.

To functionally identify the putative acyltransferases, we performed comprehensive *in vitro* functional screening for a number of the produced recombinant proteins by incubating with a collection of thioester donors and acceptors. The studies let us characterize several novel acyltransferases in relation to cell wall biogenesis and modification, which include (i) a novel monolignol acetyltransferase specifically responsible for the modification of lignin biosynthetic precursor, sinapyl alcohol. The biochemical property and subcellular localization of this enzyme were further investigated. The correlation of gene expression to lignin acetylation was explored and the transgenic plants were created. (ii) two hydroxycinnamoyltransferases that display activities for synthesis of both monolignol biosynthetic intermediates and chlorogenic acid, a metabolite acting as phytoalexin in plant defense responses. Overexpressing these genes in tobacco resulted in the interesting changes in lignin content, compositions, and xylem morphology. Additionally, we also characterized several other novel hydroxycinnamoyltransferases responsible for the modification of a variety of phenyl or aliphatic alcohols. These identified acyltransferase genes will be useful molecular tools for re-directing photosynthetic carbons from lignin polymer biosynthesis into the value-added metabolites in plant feedstocks.

USDA-DOE

Characterization of Pectin Acetylesterase Reveals Critical Roles of Cell Wall Acylesterification in Plant Growth and Development

Jin-Ying Gou* and **Chang-Jun Liu** (cliu@bnl.gov)

Biology Dept., Brookhaven National Laboratory, Upton, N.Y.

Project Goals: The goals of the project are to systemically characterize the enzymes involved in lignocellulosic acylesterification, and to understand the biochemical mechanisms and biological functions underlying cell wall modification.

Pectin is one of the structural components of plant cell wall, predominant in the primary walls of fast growing tissues. In plant, pectin is developmentally modified by methylesterification and acetylesterification. The acetylesterification of homogalacturonan usually occurs on the *O*-2 or/and *O*-3 positions in the galacturonric acid residues but its physiological function has not been well studied. In our group, a putative pectin acetylesterase homolog from poplar, ptPAE1, was cloned and heterologously expressed in *E. coli*. The recombinant protein was purified as an active enzyme form that could release acetyl moieties from the acetylated polysaccharides, pectin, xylan, and arabinogalactan. The ptPAE1 exhibits optimal activity at pH 7 and 35°C on the acetylated pectin and kinetically prefers substrate pectin over xylan. The identified esterase gene driven by a double 35S promoter was over-expressed in tobacco. The promoted de-acetylation of polysaccharides due to the constitutive expression of ptPAE1 resulted in obvious morphological phenotype. The apical buds of transgenic plants displayed constant senescence at flowering stage. In the flowers of transgenic plants, the development of filament was impeded, which resulted in short stamen and, consequently, the seed pods of the transgenic plants were greatly reduced both in size and weight, compared to the control plants. The offspring seedlings harboring transgene generated shorter hypocotyls and roots than non-transgenic control plants. The growth retardation resulted from the disturbance of pectin acetylation in the fast elongation tissues indicates that acetylesterification of pectin has significant roles in plant growth and development.

* Presenting author

USDA-DOE

Brachypodium Transcriptomics

Henry D. Priest[1]* (priesth@onid.orst.edu), Samuel Fox,[1] Wenqin Wang,[2] Todd P. Michael,[2] and **Todd C. Mockler**[1] (tmockler@cgrb.oregonstate.edu)

[1]Center for Genome Research and Biocomputing and Dept. of Botany and Plant Pathology, Oregon State University, Corvallis, Ore.; and [2]Rutgers, The State University of New Jersey, Dept. of Plant Biology and Pathology, The Waksman Institute of Microbiology, Piscataway, N.J.

Project Goals: *Brachypodium distachyon* is a model grass plant with close evolutionary relationships to economically important species including the forage and turf grasses, temperate cereals, and several potential biofuel feedstock crops. We are designing an Affymetrix *Brachypodium* whole-genome microarray that will be used for whole-genome expression profiling over major developmental stages, diurnal and circadian time-courses, and stress conditions. The resulting *Brachypodium* gene expression atlas will allow us to better understand the gene regulation networks underlying traits of major importance for both the quality and quantity of biomass.

Brachypodium distachyon is a model grass plant with close evolutionary relationships to economically important species including the forage and turf grasses, temperate cereals, and several potential biofuel feedstock crops. To better understand the gene regulation networks underlying traits of major importance for both the quality and quantity of biomass, we are pursuing a hypothesis-generating approach involving whole-genome expression profiling over major developmental stages, diurnal and circadian time-courses, and stress conditions. We are designing an Affymetrix *Brachypodium* whole-genome microarray, which will become available for purchase by any researcher. In collaboration with MIPS and JGI we are currently annotating the *Brachypodium* genome and transcriptome. A portion of the microarray space will be used to tile each predicted exon and intron with multiple probes. The remainder of the genome will be tiled to represent predicted non-coding/intergenic regions. We anticipate that the array design will be completed and Affymetrix will manufacture microarrays in the first quarter of 2009. We will use these arrays to map major gene expression changes of relevance to important traits of grass crops. We have been collecting *Brachypodium* tissues representing a number of different developmental stages, environmental conditions, and stress treatments—including (a) different tissues from plants of different developmental stages (roots, shoots, leaves, developing flower spikes, developing seeds, germinating seeds, etiolated seedlings); (b) diurnal and circadian sampling in light/dark/temperature cycles; and (c) plants exposed to abiotic stresses (heat, cold, salt, dehydration, nutrient limitation, and light stress). We are currently preparing RNAs from these samples for use with the *Brachypodium* genome arrays to establish baseline expression profiles for each sample. We will use these data to create a comprehensive public gene expression atlas. We will analyze and compare the expression patterns of *Brachypodium* genes with respect to their inferred biological functions. We will cluster co-expressed genes and perform promoter analyses to predict transcription factor binding sites in their upstream regions. Our results will be integrated into the gene annotations available through BrachyBase.org and other annotation databases, and we will develop bioinformatic tools for public access to these resources.

USDA-DOE

Identification of Cell Wall Synthesis Regulatory Genes Controlling Biomass Characteristics and Yield in Rice (*Oryza sativa*)

Zhaohua Peng[1]* (zp7@BCH.msstate.edu), Pamela Ronald,[2] and Guo-liang Wang[3]

[1]Dept. of Biochemistry and Molecular Biology, Mississippi State University, Miss.; [2]Dept. of Plant Pathology, University of California, Davis; and [3]Dept. of Plant Pathology, Ohio State University, Columbus, Ohio

Understanding the regulatory mechanisms of cell wall synthesis is essential for the improvement of biomass characteristics critical to biofuel production. To study the cell wall synthesis regulation, we are examining the cell wall regeneration process in protoplasts of rice (*Oryza sativa*). In protoplasts, the cell wall synthesis is highly activated due to the removal of cell wall. We are analyzing the transcriptome and proteome dynamic changes in response to the removal of cell wall using DNA oligo array and shotgun proteomics, respectively. The differentially regulated genes and proteins are further analyzed using systems biology method to elucidate possible pathways involved in cell wall synthesis and regeneration. In addition, mutants and transgenic lines of putative regulatory genes and critical metabolic pathway genes have been generated. Some of the mutants/transgenic lines have displayed changes in cellulose, hemicelluloses and lignin contents.

USDA-DOE

NIRS Prediction of Corn Stover Cell Wall Composition and Conversion Potential, and Relationships among these Traits

H.G. Jung[1,2]* (Hans.Jung@ars.usda.gov), M.L. Lewis,[2] R.E. Lorenzana,[2] and **R. Bernardo**[2] (bernardo@umn.edu)

[1]USDA-Agricultural Research Service, Plant Science Research Unit, St. Paul, Minn.; and [2]Dept. of Agronomy and Plant Genetics, University of Minnesota, St. Paul, Minn.

http://agronomy.cfans.umn.edu/corn_cellulosic_ethanol.html

Project Goals: 1) Determine the prospects of and identify challenges in marker-assisted breeding for both corn grain yield and stover quality traits important for ethanol production. 2) Determine if genome-wide selection (which does not require finding markers with significant effects) is superior to the usual approach of selecting only for significant markers, with the goal of simultaneously improving corn grain yield and stover quality.

Implementation of a breeding program (conventional or marker assisted) for corn stover quality traits related to cellulosic ethanol production is dependent on having a reliable, accurate, rapid, and inexpensive system for phenotyping plant material. Determination of cell wall composition and testing for fermentable sugar release in a cellulosic ethanol conversion process by conventional wet chemistry methods is laborious and expensive. However, the feed grain and forage industries have demonstrated that near-infrared reflectance spectroscopy (NIRS) is a routine and reliable system for predicting quality parameters rapidly and at low cost. We report here on the NIRS prediction equations we have developed for phenotyping corn stover in support of our project on marker-assisted selection to improve corn stover as a cellulosic ethanol feedstock. Also, correlations among the quality traits were examined for prediction of conversion potential of corn stover to determine if information on cell wall composition would be sufficient for selecting corn genotypes that have high conversion potential.

Testcrosses of 223 recombinant inbred lines (RIL) derived from B73 x Mo17 and the two parental inbreds to a tester inbred were grown at four locations in Minnesota in 2007. Corn stover, minus the ear (grain, cob, husk, and shank), was collected at grain maturity. Stover was chopped, dried at 60°C, and ground prior to scanning by NIRS (1100 to 2500 nm). A calibration sample set (N=154) was selected from the 900 total samples based on a combination of spectral diversity and physical design of the experiment. Calibration samples were analyzed for cell wall composition by the two-stage acid hydrolysis method of Theander et al. (1995) and conversion potential in a dilute acid/high temperature pretreatment and enzymatic saccharification process (Dien et al., 2006). Percent cell wall sugars released by the conversion process were determined as the difference between stover composition and the wall sugars remaining in the insoluble residue from the conversion process. NIRS calibration equations for stover quality were developed by modified partial least squares regression. Stover quality was predicted for all samples from their NIRS spectra. Mean data for corn lines were used for correlation analysis of relationships among the cell wall components and with sugar release.

The NIRS calibration equation statistics indicated that stover cell wall components can be predicted with varying precision. Glucose (primarily from cellulose) was well predicted but the other major components gave only moderate fits in the calibrations. Better calibrations have been observed for cell wall composition of corn plant part, alfalfa stem, and switchgrass herbage sample sets; however, those studies included more variability in composition due to multiple maturity stages which generally improves NIRS calibration statistics. It was observed that calibration of glucose release by the conversion process had a high R^2 whereas release of other cell wall sugars gave poor calibrations. In contrast, *in vitro* rumen digestibility of all cell wall sugars have high calibration statistics. This difference can be explained by the fact that glucose release in the conversion process and rumen digestibility are both enzymatic (biochemical) processes and cell wall matrix structure interacts with enzyme accessibility to polysaccharides. This is unlike hydrolysis of non-glucan polysaccharides in the conversion process which was primarily an acid (chemical) reaction that should be less influenced by wall matrix structure. NIRS measures organic matrix structure of biomass in the form of hydrogen bond stretching, not actual degradation of polysaccharides by either chemical or biochemical reactions.

Although NIRS calibration statistics were variable among the cell wall traits, we were able to determine that the corn RILs contained significant genetic variation for all traits and identify important quantitative trait loci. As expected for a biochemical process, release of glucose was negatively correlated with lignin concentration (r = -0.67), similar to what has been observed for rumen digestibility. Xylose and other non-glucan sugars were correlated with glucose release to varying degrees (r = -0.13 to 0.56), but were also correlated with lignin concentration. These cell wall sugar correlations with glucose release are most likely a reflection of extent of cell wall development, rather than an impact on glucose release *per se*. Glucose concentration was not related to glucose release. Release of xylose and arabinose were not correlated with lignin concentration, and uronic acid release was only poorly correlated with lignin (r = -0.25). Again, because non-glucan sugar release was a chemical hydrolysis process we did not expect large effects of lignin on their release by this conversion process. It appears that release of glucose, but not other wall sugars, can be effectively predicted from cell wall composition.

NIRS calibration statistics for corn stover quality traits and the observed range for predicted data.

Trait	Samples Used	Mean	SEC	R^2	Range
		g kg-1 DM			g kg-1 DM
Klason lignin	137	167	6	0.66	161 - 174
Glucose	135	344	5	0.92	322 - 361
Xylose	138	203	8	0.46	194 - 212
Arabinose	141	27	1	0.67	26 - 29
Uronic acids	138	27	1	0.50	26 - 28
		%			%
Glucose release	138	51.8	1.6	0.83	48.0 – 55.5
Xylose release	142	85.0	1.4	0.12	84.6 – 85.6
Arabinose release	136	88.1	1.2	0.30	87.4 – 89.0
Uronic acid release	137	75.4	2.3	0.12	74.9 – 76.4

References

1. Dien, B.S., H.G. Jung, K.P. Vogel, M.D. Casler, J.F.S. Lamb, L. Iten, R.B. Mitchell, and G. Sarath. 2006. Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. *Biomass Bioenergy* 30:880-891.
2. Theander, O., P. Aman, E. Westerlund, R. Andersson, and D. Pettersson. 1995. Total dietary fiber determined as neutral sugar residues, uronic acid residues, and Klason lignin (The Uppsala Method): collaborative study. *J. AOAC Int.* 78:1030-1044.

USDA-DOE

Epigenomics of Development in *Populus*

Kyle R. Pomraning,[2] Palitha D. Dharmawardhana,[3] Henry D. Priest,[4] Olga Shevchenko,[3] Caiping Ma,[3] Elizabeth Etherington,[3] Larry Wilhem,[4] Todd Mockler,[1,4] Michael Freitag,[1,2] and **Steven Strauss**[1,3]*

[1]Center for Genome Research and Biocomputing, [2]Dept. of Biochemistry and Biophysics, [3]Dept. of Forest Ecosystems and Society, and [4]Dept. of Botany and Plant Pathology, Oregon State University, Corvallis, Ore.

Project Goals: Epigenetics is defined by long-lasting or heritable changes in gene expression that are not associated with changes in DNA sequence. It is mainly reflected in methylation of DNA and chemical changes in DNA-associated chromosomal proteins such as histones. Recognition of its importance as a means for control of plant development has increased significantly in recent years, however, little is known about epigenetic controls in the life of trees and other woody plants. Many traits important to biomass growth and adaptability in trees may be under epigenetic control, thus may be useful for their breeding and biotechnology. This includes timing of flowering and flower structure; dormancy induction and release; shoot and leaf architecture; amenability to organ regeneration; stress tolerance; and phase-associated changes in wood structure. We will use poplar (genus *Populus*, including aspens and cottonwoods), because it has been designated as a model woody biomass species for genomic studies, and is a major source of wood, energy, and environmental services in the USA and throughout the world. We will characterize epigenetic changes in DNA methylation and two kinds of histone modification via a combination of antibody-based chromatin immunoprecipitation and DNA sequencing ("ChIP-sequencing").

Poplar (genus *Populus*) is a deciduous tree that provides a variety of environmental services and is used in many countries for agroforestry, bioenergy, pulp and wood products. Poplar is widely considered the model taxon for woody plant genomics and biotechnology. To better understand how chromatin structure varies during tree development, we are determining the distribution and variation of cytosine DNA methylation and specific histone modifications (1) in different poplar tissues, (2) during the annual dormancy cycle, (3) during in vitro regeneration, and (4) after transgenic suppression of DNA methylation by RNAi-mediated silencing of the two predicted poplar ddm1 homologues. We collected root, leaf, bud, xylem, phloem, and callus samples, and subjected them to methylated-DNA immunoprecipitation (MeDIP) and chromatin immunoprecipitations (ChIP) followed by high throughput Illumina/Solexa sequencing. By using antibodies specific against 5-methylcytidine, trimethylated histone H3 lysine 4 (H3 $K4^{Me3}$), and H3 $K9^{Me3}$ and H3 $K27^{Me3}$, we will determine the genome-wide distribution of methylated DNA and the selected H3 marks. Except for H3 $K4^{Me3}$ all of these modifications are considered heterochromatic (silencing) marks. As expected, MeDIP and quantitative PCR studies of leaf chromatin have revealed high levels of cytosine DNA methylation at a variety of predicted transposable elements, and low levels of cytosine methylation at predicted active regions with high transcriptional activity. We are currently optimizing ChIP conditions and in vitro library construction for sequencing, and have begun to collect high-throughput DNA sequencing data. Epigenomic data are analyzed by a custom-developed, efficient pipeline for Solexa data analysis and by graphical representation on a poplar genome browser.

USDA-DOE

An EST-Microsatellite Linkage Map of Switchgrass (*Panicum Virgatum* L.) and Comparison Within the Poaceae

Christian Tobias[1] (christian.tobias@ars.usda.gov), Miki Okada[1]* (miki.okada@ars.usda.gov), Christina Lanzatella-Craig,[1] Brindha Narasimhamoorthy,[2] Malay Saha,[2] and Joe Bouton[2]

[1]USDA-ARS Western Regional Research Center, Calif.; and [2]Samuel Roberts Noble Foundation, Inc., Forage Improvement Division, Okla.

Project Goals: Create a gene based linkage map of tetraploid switchgrass for marker-trait association.

Switchgrass is widely viewed as a promising crop for bioenergy production. However, development of improved cultivars optimized for bioenergy through breeding involves improving yields and altering feedstock composition so that competition for limited arable land is minimized and process efficiencies are fully realized. Fundamental to any advanced breeding program are availability of molecular markers and genetic linkage maps that facilitate modern cultivar development through marker assisted selection (MAS). New crops such as switchgrass stand to benefit from the application of MAS techniques and through comparative approaches with other grasses that will provide a multitude of candidate gene-loci for traits considered important for bioenergy.

As a precondition to SNP and microsatellite development 500,000 EST were produced at the DOE Joint Genome

Institute from 10 distinct sources that included multiple genotypes and tissues. These were assembled into 74,869 consensus sequences; a large number which reflects the allelic diversity of this species. Approximately seventy percent of the assembled consensus sequences could be aligned with the sorghum genome at an *E*-value of <1 x 10^{-20} indicating a high degree of similarity. Splice junction and coding sequence predictions were produced based on these alignments. The representation in the libraries of gene families known to be associated with C4 photosynthesis, cellulose and beta-glucan synthesis, phenylpropanoid biosynthesis, and peroxidase activity indicated likely roles for individual family members. Comparisons of synonymous codon substitutions were used to assess genome sequence diversity and indicated an overall similarity between the two genome copies present in the tetraploid.

Identification of EST-SSR markers and amplification on two individual parents of a tetraploid F1 mapping population yielded an average of 2.18 amplicons per individual and 35% of the markers produced useful fragment length polymorphisms. We have produced a microsatellite map from this F1 mapping population that is both an initial step toward QTL mapping and part of a larger project to produce an integrated genetic map that incorporates genic and intergenic markers, determines the extent of preferential pairing/disomic inheritance in switchgrass and delimits the two distinct genomes. Mapping of an advanced F2 generation which will facilitate linkage phase and haplotype determination has also been initiated. This F2 population is being established in a replicated field trail for analysis of biomass characteristics and mapping of quantitative trait loci. Thus far a total of 522 markers at 308 loci have been mapped in both parents with map lengths of 1947 cm and 2345 cM for each parent. Currently we detect 38 linkage groups organized into 9 homeologous groups. However we expect this number to stabilize at 36 when genomic SSR and EST-SSR datasets are integrated. Extensive collinearity with sorghum is apparent with the differences in base chromosome number appearing to result from the fusion of sorghum chromosomes 8 and 9.

USDA-DOE

Fast Detection of Improved Hydrolysis in Plants with Genetically Modified Lignin

C. Chapple[1] (chapple@purdue.edu), R. Meilan,[2] X. Li,[1] E.A. Ximenes,[3,4] Y. Kim,[3,4] and M.R. Ladisch[3,4]*

[1]Dept. of Biochemistry, [2]Dept. of Forestry and Natural Resources, [3]Dept. of Agricultural and Biological Engineering, and [4]Laboratory of Renewable Resources Engineering, Potter Engineering Center, Purdue University, West Lafayette, Ind.

Project Goals: The phenolic cell wall polymer lignin constitutes a significant barrier to biomass conversion but, at the same time, is essential to normal plant growth and development. This research addresses efforts to rationally engineer lignification in species used as energy crops in a manner that will help to maximize biomass-to-energy conversion without impacting the vital functions of lignin. The work is utilizing molecular strategies and appropriate transgenes for manipulating the expression of lignification-associated genes (Chapple lab); generating and evaluating poplar engineered for altered lignin content and/or monomer composition, and field-testing this perennial woody plant for fitness (Meilan lab); analyzing the impact of these transgenic strategies on metabolism in general and lignin biosynthesis in particular (Chapple lab); and evaluating the ease with which cell wall deconstruction can be accomplished using both chemical and enzymatic means (Ladisch lab). The research has four major objectives: 1) Generation of transgenic poplar up- or down-regulated for four enzymes known to impact lignin quantity and quality; 2) Development of metabolic profiling methods for poplar and their application to greenhouse- and field-grown wild-type and transgenic plants; 3) Field-plot morphometric analysis of transgenic lines; and 4) Cell wall deconstruction analysis of wild-type and lignin-modified transgenic lines.

The phenol cell wall polymer, lignin, constitutes a significant barrier to biomass conversion, but at the same time is essential to normal plant growth and development. Plant genetic engineering has the potential to reduce costs for biofuel production by enabling modifications in plant cell wall structure and lignin to acheive higher digestibility of the cellulose and reducing the need for pretreatment. Pretreatment disrupts the cell wall structure so that cellulase enzymes may readily access the cellulose and break it down to glucose. This research addresses efforts to rationally engineer lignification in hybrid popular (species *Populus*) in order to minimize the severity of the pretreatment and thereby reduce the cost and enhance the transformation of structural polysaccharides in plant cells to fermentable sugars.

The ability to control cell-wall composition without compromising plant performance is a challenge and key objective of bioenergy crop improvement. Identification of successful plant modifications entails metabolic profiling methods applied to greenhouse- and field-grown wild-type and transgenic plants. While profiling gives insights on effects of the up- or down-regulation of the four enzymes known to impact lignin synthesis, comparison of changes in genetics to changes in bioprocessing properties that enhance the production of ethanol is needed. A method for the rapid detection of biodegradability in genetically modified plants (*Arabidopsis thaliana* and *Populus*) that vary in lignin content and/or composition is addressed as part of this research. Pretreatment is carried out by pressure cooking 50 mg cellulosic material in water at 200 C for 10 min. The samples are then cooled, enzyme is added, and hydrolysis carried out for 30 min at 50 C and pH 4.8. The liquid is assayed for the amount of glucose formed. The method enables the quick evaluation of individual plants, grown in the greenhouse, since each plant weighs 50 to 70 mg, and hence one plant is equivalent to one sample. Using the reported method, we have been able to rapidly and reproducibly identify genetically modified plants that have improved biodegradability.

Extended hydrolysis times of 48 to 72 hours for *Populus* have resulted in conversions approaching 80%. These methods are identifying both *Arabidopsis* and *Populus* plants for use in further developing lignocellulosic feedstocks for biofuels production.

USDA-DOE

Genome-Enabled Discovery of Carbon Sequestration Genes

Gerald A. Tuskan[1] (tuskanga@ornl.gov), Timothy J. Tschaplinski[1]* (tschaplinstj@ornl.gov), Udaya Kalluri,[1] Tongming Yin,[1] Xiaohan Yang,[1] Xinye Zhang,[1] Nancy Engle,[1] Priya Ranjan,[1] Manojit Basu,[1] Lee Gunter,[1] Sara Jawdy,[1] Madhavi Martin,[1] Alina Campbel,[1] Stephen DiFazio,[2] John Davis,[3] Maud Hinchee,[8] Christa Pinnacchio,[6] Rick Meila,[4] Victor Buso,[7] and Steve Strauss[5]

[1]Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]University of West Virginia, Morgantown, W. Va.; [3]University of Florida, Gainesville, Fla.; [4]Purdue University, West Lafayette, Ind.; [5]Oregon State University, Corvallis, Ore.; [6]DOE Joint Genome Institute, Walnut Creek, Calif.; [7]Michigan Technological University, Houghton, Mich.; and [8]ArborGen LLC, Summerville, S.C.

Background and goals - The fate of carbon below ground is likely to be a major factor determining the success of carbon sequestration strategies involving plants. Despite their importance, molecular processes controlling belowground C allocation and partitioning are poorly understood. This project is leveraging the *Populus trichocarpa* genome sequence to discover genes important to C sequestration in plants and soils. The focus is on the identification of genes that provide key control points for the flow and chemical transformations of carbon in roots, concentrating on genes that control the synthesis of chemical forms of carbon that result in slower turnover rates of soil organic matter (i.e., increased recalcitrance). We propose to enhance carbon allocation and partitioning to roots by 1) modifying the auxin signaling pathway, and the invertase family, which controls sucrose metabolism, and by 2) increasing root proliferation through transgenesis with genes known to control fine root proliferation (e.g., ANT), 3) increasing the production of recalcitrant C metabolites by identifying genes controlling secondary C metabolism by a major mQTL-based gene discovery effort, and 4) increasing aboveground productivity by enhancing drought tolerance to achieve maximum C sequestration. This broad, integrated approach is aimed at ultimately enhancing root biomass as well as root detritus longevity, providing the best prospects for significant enhancement of belowground C sequestration.

Transgenic poplar lines with altered root phenotypes - Phenotypic analysis of root development in Aux7, Aux17, Aux16-1, 717-1B4, and transformation control were carried out as these lines had a visual aboveground phenotype. By week 8, the number of primary and lateral roots was significantly lower in Aux7 than any other line. A number of activation-tagged poplar clones with altered rooting were also identified. Those with increased rooting included D24-4, K694-2-1, D23-12, D23-12, D16-20, D18-29, and D18-16. Clones with decreased rooting included D18-84, D18-49, D23-2, E4-9, and E18-14.

Metabolic profile of activation-tagged poplar mutants - The metabolic profiling of activation-tagged poplar clones (roots and leaves) that displayed extremes (increase and decrease) in rooting response was completed. The profiles range from very few differences between the mutant and wild type control to a large number of extensive changes. Three of the clones with reduced rooting had a reduced concentration myo-inositol in roots. Myo-inositol was shown to be highly correlated with rooting in poplar stem cuttings (Tschaplinski and Blake 1989). The clones with increased rooting tended to have increases in aspartic acid and other amino acids in the roots, as seen in ANT mutants with increased rooting. This mutant also had large accumulations of complex phenolic glycosides at the expense of lower MW phenolics. Although some responses are common among the mutants, each mutant needs to be considered separately. The bases for the underlying changes will be elucidated by recovering the 35S promoter region using TAIL-PCR and sequencing in both directions to identify the genes that may be directly or indirectly contributing to the observed profiles.

Metabolite QTL and underlying candidate genes - Several secondary carbon metabolite quantitative trait loci (mQTL) were identified in a hotspot on poplar LG X. The first five candidate genes underlying the identified mQTL were selected. Constructs for up and down regulation were created and submitted to ArborGen for transformation into *P. deltoides*. We anticipate receiving these transformants early next year.

Productivity QTL - A primary focus of this project has been the examination of the genetic control of biomass allocation above and below ground. We examined controls of productivity in TDxD' Family 52124 containing over 800 hybrid clones. We developed a hydroponic system that provides robust growth and used this system to assess biomass allocation in 252 genotypes from this pedigree using two replicates in a completely randomized design. The clonal repeatability, or genotypic broad-sense heritability, ranged from 0.286 to 0.333 for three biomass-allocation traits, indicating that the phenotypes are repeatable between replicates. These traits are currently being mapped as QTL, utilizing a genetic map developed for this pedigree.

Poplar drought tolerance - Cisgenic *Populus deltoides* clones were down- and up-regulated with a constitutive promoter with drought tolerance genes identified in QTL intervals. The genes selected were all known transcription factors that were previously identified in other species. An up-regulated popMYB clone was 16% more productive than the wild-type control and 54% more productive than the empty vector control. A number of the transgenic lines had unusual growth forms, exhibiting a loss of erect growth habit as a result of the constitutive promoter. A second round of trans-

genic plants was created using a drought-inducible promoter identified from qRT-PCR analyses as highly responsive in a controlled drought experiment, increasing 27238-fold in a *P. deltoides* clone under severe drought stress. A global transcript profile is being conducted in *P. deltoides* that will identify novel drought response pathways. Enhanced drought tolerance will ensure sustained productivity and maximum C sequestration potential.

This research was supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

USDA-DOE

Genetic Dissection of Bioenergy Traits in Sorghum

Wilfred Vermerris[1]* (wev@ufl.edu), Ana Saballos,[1] Zhanguo Xin,[2] Jeffery Pedersen,[3] Scott Sattler,[3] William Rooney,[4] Seth Murray,[4] and Stephen Kresovich[5]

[1]University of Florida, Gainesville, Fla.; [2]USDA-ARS, Lubbock, Tex.; [3]USDA-ARS, Lincoln, Nebr.; [4]Texas A&M University, College Station, Tex.; and [5]Cornell University, Ithaca, N.Y.

Project Goals: The overall goal of the project is to identify the genetic basis of a number of traits that enhance the value of sorghum as a bioenergy crop. These traits relate to the production of fermentable sugars from the sugar-rich juice in the stems of sweet sorghums and from the cell wall polysaccharides in sorghum stover or bagasse. The specific objectives are to: 1) identify the gene(s) underlying a major QTL for stem sugar concentration, 2) identify QTL for stem juice volume and stalk sugar concentration, 3) classify approximately 60 novel sorghum brown midrib (bmr) mutants from the USDA TILLING population in allelic groups based on cell wall chemistry and allelism tests, 4) select representative bmr mutants from each allelic group for their potential value as feedstock for ethanol production, and 5) clone and characterize those Bmr genes that represent loci other than the two cloned Bmr genes, using a mapping and a candidate gene approach.

Sorghum (*Sorghum bicolor* (L.) Moench) is a multi-source bioenergy crop that can produce grain, sugar-rich juice and lignocellulosic biomass as feedstocks for fermentable sugars for the production of liquid transportation fuels and green chemical feedstocks. Alternatively, production schemes in which some of the products are used for other purposes, including animal feed or biomass for the generation of heat or electricity are possible. Sorghum requires lower inputs than maize and sugar cane and is tolerant of a wide range in temperature and soil composition.

We are investigating the genetic basis of a number of traits in order to enhance their utility and facilitate plant breeding efforts. The *brown midrib* (*bmr*) mutations affect the chemical composition of the cell wall. Some of the *bmr* mutations have been shown to result in at least a 50% increase in the yields of glucose and xylose after thermo-chemical pretreatment and enzymatic saccharification of sorghum stover. There are at least four *Bmr* genes, two of which have been cloned and shown to encode enzymes in the monolignol biosynthetic pathway. An EMS-mutagenized population containing close to 60 *bmr* mutants generated by the USDA-ARS is being used to identify additional *bmr* loci and novel alleles of already established loci, and to establish the function of the *Bmr* genes that have not yet been cloned.

The stems of sweet sorghums contain sugar-rich juice that can be used directly for fermentation. Total sugar yield is determined by juice volume and sugar concentration. We are using a recombinant inbred line (RIL) population derived from a cross between a sweet sorghum and a dry-stalk, non-sweet sorghum to map quantitative trait loci (QTL) involved in determining juice volume. Prior research using a different RIL population led to the identification of a major QTL for sugar concentration in sweet sorghum. We are using genetically similar lines with contrasting genotypes for this QTL derived from this population as a basis for high-throughput expression profiling based on the 454 sequencing platform to identify the gene(s) in this QTL.

Our combined efforts are expected to yield novel insights in the genetic control of primary metabolism in sorghum, and result in genetic markers that will greatly expedite the development of the next generation of bioenergy sorghums.

USDA-DOE

Insertional Mutagenesis of *Brachypodium distachyon*

John Vogel* (john.vogel@ars.usda.gov), Jennifer Bragg, Jiajie Wu, Yong Gu, Gerard Lazo, and Olin Anderson

USDA-ARS, Western Regional Research Center, Albany, Calif.

Project Goals: 1) Generate >7,500 insertional mutants in the model grass *Brachypodium distachyon*; 2) Sequence DNA flanking >6,000 insertion sites and annotate the genes affected using the genomic sequence; and 3) Establish a website where researchers could search the flanking sequence database and order knockouts in genes of interest.

Brachypodium distachyon (*Brachypodium*) is an excellent model for studying the basic biological processes underlying the traits that determine the utility of grasses as energy crops. Several important genomic resources have been developed in *Brachypodium*, including the recently completed whole genome sequence, ESTs, a high-density genetic linkage map and germplasm resources. The objective of our work is to generate >7,500 insertional mutants and to sequence the regions flanking >6,000 insertion sites. The location of insertions in the genome will be determined

* Presenting author

by comparing flanking sequences to the complete genome sequence. This information will be loaded into a searchable website to provide researchers with a means to order T-DNA lines with mutations in genes of interest. Thus, this project will provide a large, freely available collection of sequence-indexed mutants to researchers studying grasses and grains. This poster describes our experiments aimed at optimizing transformation, evaluating transposon tagging systems and sequencing flanking DNA. As a first step toward designing constructs for large scale mutagenesis, we compared several vectors and observed considerable variation in efficiency. The average transformation efficiency (from 6 independent transformations) of the best construct was 55% with a mean survival rate of 88% after the plants were transferred to soil. We improved our transformation method by modifying the callus initiation media and omitting a recovery step. These changes increased transformation efficiency, decreased labor by eliminating a transfer step and decreased the time needed to create transgenics. To date, we have generated >1,300 T_0 T-DNA mutant lines. Over 95% of T_0 plants tested showed expression of a GUS reporter gene, and all were positive by PCR for the presence of the selectable marker. T_1 seed has been harvested for 450 lines and 225 lines have been planted for phenotypic evaluation. To determine if transposon tagging is more efficient that T-DNA tagging we evaluated both Ac/Ds and En/Spm transposon systems in *Brachypodium*. Both transposons were lethal to the majority of transgenic plants indicating that neither transposon system is compatible with *Brachypodium*. We are currently comparing TAIL-PCR and inverse-PCR to sequence flanking DNA. Initial experiments suggest that TAIL-PCR is less efficient than inverse-PCR. To date, we have generated 79 flanking sequences from 121 T_0 plants using inverse-PCR. Of these sequences, 45 (57%) contain *Brachypodium* genomic sequence, and 34 (43%) contained only vector sequence. We are currently transitioning from optimization to production scale generation of T-DNA lines.

USDA-DOE

Identification of Genes That Control Biomass Production Using Rice as a Model System

Jan E. Leach[1]* (jan.leach@colostate.edu), Daniel R. Bush[2,3] (dbush@colostate.edu), Hei Leung[3] (h.leung@cgiar.org), and John McKay[1,3] (jkmckay@lamar.colostate.edu)

[1]Bioagricultural Sciences and Pest Management and [2]Department of Biology, Colorado State University, Ft. Collins, Colo.; and [3]International Rice Research Institute, Manila, Philippines

Project Goals: Our goal is to provide the applied biomass research community and industry with information to allow exploitation of the genes and pathways relevant to biomass accumulation in grasses.

Phenotypic variation in cellulosic biomass accumulation has not been systematically investigated in any plant species. We propose a critical first step to identifying the genotypic variation that controls biomass productivity, using rice as a model grass. Using the genetic and genomic tools available in rice, we aim to establish causal links between the genes (or genetic regions in the case of multigenic traits), gene expression profiles, and phenotype (biomass accumulation). Identifying the loci that are essential for optimal cellulosic biomass in rice will provide a set of loci and networks that are molecular targets for improving target biomass crops (e.g., *Miscanthus*, switchgrass). We have identified three lines of rice that span the variation in biomass accumulation and for which we have comprehensive SNP data. One population has already been advanced to 700 F6 recombinant inbred lines (RILs); development of the second population is in progress. These genetic resources will be used to dissect the genetic basis of biomass accumulation.

Using these two populations, we will (1) identify QTL for biomass accumulation phenotypes, (2) refine these QTLs using an integrated analysis of variation at both the DNA and expression levels as well as analysis of mutant populations, and (3) validate gene candidates for biomass traits using mutant analysis, gene silencing strategies or overexpression analysis. Detailed physiological and morphological descriptions of the parents, high-biomass RILs and mutants, and transgenic plants used in validation experiments will parallel genetic analysis. Gene expression profiles will be generated from near-isogenic lines (NILs) at two growth stages in a single season to identify genomic regions where gene expression of clustered genes is correlated with phenotype. All data will be integrated into a convenient gene browser to facilitate identification of gene candidates for the biomass traits. Finally, we will validate the significance of a suite candidate genes or pathways in biomass accumulation by mutant analysis, gene silencing (RNAi), and/or overexpression. Screening and field assessments as well as data integration will involve exchange of PIs and postdoctoral fellows between CSU and IRRI. Deliverables from this project include: (1) Characterized genomic regions in rice that are associated with variation relevant to biomass production, (2) Permanent rice genetic stocks (e.g., RILs and NILs) that will enable further exploitation of biomass diversity and other desirable traits, (3) selected candidate genes associated with biomass production, (4) validation of function of genes predicted to be involved in biomass production.

In the first few months of funding, we have completed a comprehensive phenotyping for biomass traits (plant size and architecture, photosynthetic rates and assimilate partitioning) for the three rice lines that serve as parents for the RIL populations, and based on these results, have initiated phenotypic screening of the available RIL population. Development of the second RIL population is underway. A comprehensive SNP dataset and phenotypic and genetic mapping information are being used to associate genotype and phenotype and identify candidate biomass genes. Over 15 putative mutants altered in biomass accumulation (large plants) have been identified through phenotypic screens of

a deletion collection. After confirmation of genotype, DNA from these mutants will be hybridized to rice oligonucleotide arrays for discovery of deleted regions.

USDA-DOE

Resource Development in Switchgrass, a Polyploid Perennial Biofuel Grass

Katrien M. Devos[1,2*] (kdevos@uga.edu), Jeffrey L. Bennetzen,[3] Joe H. Bouton,[4] E. Charles Brummer,[1] and Malay C. Saha[4]

[1]Institute of Plant Breeding, Genetics, and Genomics, Dept. of Crop and Soil Sciences, [2]Dept. of Plant Biology, and [3]Dept. of Genetics, University of Georgia, Athens, Ga.; and [4]Forage Improvement Division, The Samuel Roberts Noble Foundation, Ardmore, Okla.

Project Goals: This project aims to develop tools such as DNA markers, genetic maps and trait maps to enhance the breeding of switchgrass for bioenergy production. We are also interested in gaining insight in the structural organization of the switchgrass genome, and understanding the relationship of the switchgrass genome with that of the model species foxtail millet as a foundation for the future sequencing of the complex switchgrass genome.

Switchgrass, *Panicum virgatum*, is a perennial grass that is native to the United States. It is a polyploid that has been bred as a forage crop for more than a decade. Its high biomass potential under low inputs also makes switchgrass a good candidate biofuel crop. The efforts to develop switchgrass as a dedicated biofuel crop have gained impetus in the past few years with the funding of several collaborative projects to develop genetic tools and populations for trait mapping. Our project is focused on the development of resources for switchgrass, including SSR markers, genetic maps, trait maps, a fosmid library and a limited amount of sequence information to get insight into the structural organization of the switchgrass genome. Mapping will be carried out in a biparental population developed from a cross between a lowland Alamo genotype and an upland Summer genotype. This population has been planted in two locations and a first round of phenotyping has been carried out in the fall of 2008. Genetic mapping is in progress. Markers for mapping include EST-derived and genomic simple sequence repeats (SSRs), candidate genes for biomass yield and, potentially, single nucleotide polymorphism (SNP) markers. The genetic maps will be aligned with the maps of other grass species, in particular foxtail millet. Foxtail millet is a small-genome relative of switchgrass whose whole genome sequence is expected to become available in Spring 2009 and will act as a model for switchgrass genomics. The switchgrass and comparative tools will greatly accelerate the breeding of switchgrass varieties targeted at enhanced bioenergy conversion potential.

USDA-DOE

Identification of Genes that Regulate Phosphate Acquisition and Plant Performance During Arbuscular Mycorrhizal Symbiosis in *Medicago truncatula* and *Brachypodium distachyon*

S. Karen Gomez,[1] Nathan Pumplin,[1] Aynur Cakmak,[1] Jeon Hong,[1] Ying Li,[2] Matthew E. Hudson,[2] and **Maria J. Harrison**[1*] (mjh78@cornell.edu)

[1]Boyce Thompson Institute for Plant Research, Ithaca, N.Y.; and [2]Dept. of Crop Sciences, University of Illinois, Urbana, Ill.

Project Goals: The overall goals are to identify and characterize plant genes involved in the regulation and functioning of the AM symbiosis in *Medicago truncatula* and *Brachypodium distachyon*. The specific objectives are: 1) Identify and characterize *M. truncatula* transcription factors expressed in mycorrhizal roots. 2) Identify transcription factors required for development of the arbuscule-cortical cell interface and symbiotic phosphate transport. 3) Investigate the molecular basis of functional diversity among AM symbioses in *B. distachyon* and identify gene expression patterns associated with AM symbioses that differ functionally with respect to plant performance. 4) Analyze the predicted *B. distachyon* ortholog of MtPT4 and its role in symbiotic phosphate transport.

Most vascular flowering plants are able to form symbiotic associations with arbuscular mycorrhizal (AM) fungi. The symbiosis develops in the roots and has a profound effect on plant productivity, largely through improvements in plant mineral nutrition (1). All proposed bioenergy crops including legumes, grasses and trees, are capable of forming AM symbioses, and therefore have the potential to benefit from phosphorus and nitrogen acquisition through the symbiosis. This is significant because phosphorous and nitrogen are the two mineral nutrients whose availability is most frequently limiting for plant growth (2).

To ensure that future crops contain the optimal set of alleles to benefit maximally from an AM symbiosis, it is necessary first to understand which genes control the AM symbiosis and the molecular basis of their function. To address this we aim to identify plant transcription factors that regulate development of the AM symbiosis and symbiotic phosphate transfer in *Medicago truncatula*. In mycorrhizal roots, symbiotic phosphate transfer occurs in the cortical cells harboring fungal arbusucles and these cells comprise only a small proportion of the total cortical cell population within the root. To identify transcription factors operating in this cell type, we have developed laser capture microdissection to enable the capture of cells containing arbuscules. Analysis of transcription factor gene expression in this cell type is in progress. In parallel, we are taking advantage of genomics resources developed in M. truncatula to enable bioinformatics approaches to identify potential cis-regulatory motifs

associated with promoters of AM symbiosis-induced genes (3).

M. truncatula is widely used as a model for studies of AM symbiosis but it is not an ideal model for analysis of AM symbiosis in grass species. *Brachypodium distachyon* is a wild grass species that serves as a model for the temperate grasses, including those proposed as bioenergy crops. There are several detailed ecological studies of AM symbioses of grasses, including studies of the genus *Brachypodium* (4,5), and these indicate the potential of the AM symbiosis for increasing plant growth in low phosphate soils. However, these studies also illustrate significant differences in plant performance depending on the AM fungal species involved. Variation in plant performance during symbiosis with different AM fungal symbionts is a well documented phenomenon; however, the molecular basis is not understood. We aim to use *B. distachyon* to investigate this and will use sequence-based transcript profiling and statistical analyses to compare transcript profiles in *B. distachyon* AM symbioses that differ in their effect on plant performance. The goal is to identify transcript profiles and potentially plant processes associated with maximal plant performance. Experiments to evaluate the response of *B. distachyon* to different AM fungi are in progress.

References

1. Javot H, Penmetsa RV, Terzaghi N, Cook DR, Harrison MJ. 2007. A *Medicago truncatula* phosphate transporter indispensable for the arbuscular mycorrhizal symbiosis. *PNAS* 104: 1720-5
2. Vance C. 2001. Symbiotic Nitrogen Fixation and Phosphorus Acquisition. Plant Nutrition in a World of Declining Renewable Resources. *Plant Physiology* 127: 390-7
3. Hudson ME, Quail PH. 2003. Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiology* 133: 1605-16
4. van der Heijden MGA, Wiemken A, Sanders IR. 2003. Different arbuscular mycorrhizal fungi alter coexistence and resource distribution between co-occuring plant. *New Phytologist* 157: 569-78
5. van der Heijden MGA, Streitwolf-Engel R, Riedl R, Siegrist S, Neudecker A, et al. 2006. The mycorrhizal contribution to plant productivity, plant nutrition and soil structure in experimental grassland. *New Phytologist* 172: 739-52

Biofuels > Analytical and Imaging Technologies for Studying Lignocellulosic Material Degradation

GTL

Three-Dimensional Spatial Profiling of Lignocellulosic Materials by Coupling Light Scattering and Mass Spectrometry

P.W. Bohn[1]* (pbohn@nd.edu) and J.V. Sweedler[2] (sweedler@scs.uiuc.edu)

[1]Dept. of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Ind.; and [2]Dept. of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Ill.

Project Goals: The project goals are to combine Raman imaging and MALDI and SIMS mass spectrometric imaging to accomplish four main objectives: (1) develop correlated optical and mass spectrometric imaging approaches for sub-surface imaging to address the highly scattering nature of LCMs in different states of processing; (2) adapt mass spectrometric imaging protocols to enable spatially-resolved LCM characterization; (3) create surface optical and mass spectrometric measures of lignin-hemicellulose-cellulose degradation at specific processing stages; and (4) correlate the in situ optical (Raman and SH-OCT) and mass spectrometric information to generate depth-resolved maps of chemical information as a function of spatial position and processing time.

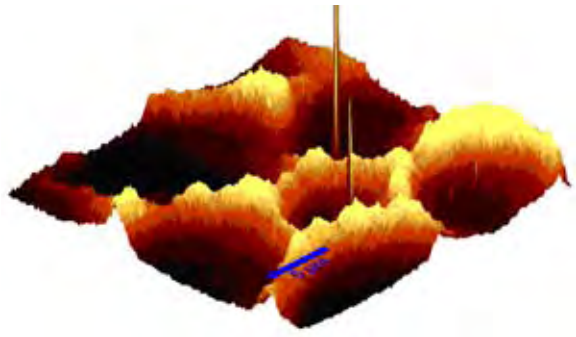


Figure. 3D rendering of lignin from untreated *Miscanthus*.

The physical and chemical characteristics of lignocellulosic materials (LCMs) pose daunting challenges for imaging and molecular characterization: they are opaque and highly scattering; their chemical composition is a spatially variegated mixture of heteropolymers; and the nature of the matrix evolves in time during processing. Any approach to imaging these materials must (1) produce real-time molecular

speciation information *in situ*; (2) extract sub-surface information during processing; and (3) follow the spatial and temporal characteristics of the molecular species in the matrix and correlate this complex profile with saccharification. To address these challenges we are implementing tightly integrated optical and mass spectrometric imaging approaches. Employing Raman microspectroscopy (RM) provides real-time *in situ* information regarding the temporal and spatial profiles of the processing species and the overall chemical degradation state of the lignin heteropolymer; while MALDI and SIMS provide spatially-resolved information on the specific molecular species produced by pre-enzymatic processing. The goals of the approach are to: (1) develop correlated optical and mass spectrometric imaging approaches for sub-surface imaging to address the highly scattering nature of LCMs in different states of processing; (2) adapt mass spectrometric imaging protocols to enable spatially-resolved LCM characterization; (3) create surface optical and mass spectrometric measures of lignin-hemicellulose-cellulose degradation at specific processing stages; and (4) correlate the *in situ* optical (Raman and SH-OCT) and mass spectrometric information to generate depth-resolved maps of chemical information as a function of spatial position and processing time. Raman and MALDI/SIMS imaging results from native (unprocessed) and acid-treated *Miscanthus x giganteus* will be described. Approaches for generating chemically specific information in plan view and three-dimensional profiling (*viz.*figure) will be also be presented.

GTL

Visualization of Acid-Pretreatment Effects on Lignocellulose by Integration of Neutron Scattering and Computer Simulation

Barbara R. Evans[1]* (evansb@ornl.gov), Jeremy C. Smith,[2,3] Loukas Petridis,[2] Volker Urban,[1,4] Sai Venkatesh Pingali,[1,4] Dean A. Myles,[1,4,5] William T. Heller,[1,4] Hugh M. O'Neill,[1,4] Joseph McGaughey,[1] Art Ragauskas,[6] Marcus Foston,[6] and Ida Lee[1,7]

[1]Chemical Sciences Division and [2]Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Governor's Chair, University of Tennessee, Knoxville, Tenn.; [4]Center for Structural Molecular Biology and [5]Neutron Scattering Science Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [6]Institute of Paper Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Ga.; and [7]Dept. of Electrical and Computer Engineering, University of Tennessee, Knoxville, Tenn.

Project Goals: The overarching goal is the integration of neutron scattering techniques, surface force imaging, and computer simulation technology to provide in situ real-time multi-scale visualization of lignocellulose structure. The comprehensive molecular and nanoscale information thus obtained will enable the targeted improvement of biomass deconstruction and saccharification that is needed for efficient cellulosic ethanol production. This new imaging technology will integrate the continuously evolving power and capabilities of neutron scattering and high-performance leadership computer simulation at Oak Ridge National Laboratory to synergistically derive information on lignocellulosic degradation at an unprecedented level of detail.

A detailed physicochemical understanding of lignocellulose structure and its deconstruction during processing at multiple length scales is needed to enable cost-effective production of fuels from biomass. To attain this goal, we are developing a real-time *in situ* multi-length scale neutron scattering-computer simulation technology with complimentary surface force recognition imaging. Key to this effort is the design and employment of multipurpose neutron imaging chambers that will enable *in situ* dynamic observation of the structural evolution of biomass under pretreatment conditions. This visualization capability resulting from the combination of neutron scattering and high-performance computer simulation afforded through facilities at ORNL will provide fundamental information about biomass morphological degradation during heat, pressure, chemical or enzymatic treatment at an unprecedented level of detail. The synergistic capabilities of the Spallation Neutron Source (SNS), the High Flux Isotope Reactor (HFIR) and the National Leadership Computing Facility (NCLF) at ORNL will be combined with supporting information from biochemistry, X-ray diffraction, small-angle x-ray scattering, and physical and chemical characterization of lignocellulosic materials through surface force recognition imaging. This novel combination of experimental and computational techniques will produce the fundamental molecular scale understanding of biomass degradation required to develop cost-effective approaches to cellulosic ethanol production.

We characterized samples of standard biomass feedstocks of loblolly pine, switch grass, and hybrid poplar provided by the Institute of Paper Science and Technology (IPST) at Georgia Tech. The IPST provided technical background on composition of the lignocellulose in the feed stock plants and samples of the component biopolymers cellulose, hemicellulose, and lignin that provided the separate scattering signatures of the components to assist in the analysis of the scattering and diffraction patterns of whole lignocellulose. Small-angle neutron scattering (SANS) experiments were carried out on the Bio-SANS beamline at the ORNL High Flux Isotope Reactor (HFIR). In the preliminary experiments, untreated biomass samples loblolly pine, switch grass, and hybrid poplar, and the individual components extracted from these samples were examined. Matched sets of the isolated components cellulose, hemicellulose, and lignin were prepared at IPST by standard methods. The scattering patterns of these components are being used to assist in analysis of the patterns obtained from the corresponding whole plant material. Additionally, the structural consequences of dilute sulfuric acid pretreatment of switch grass and poplar were studied by measuring a pretreatment reaction-time series. These results are being used to determine the parameters required for design and use of the real-time visualization

* Presenting author

chambers. The results suggest that very little changes in the sample over the length scales investigated during this standard pretreatment process. A force field for molecular dynamics of cellulose has been developed (Petridis and Smith, *J. Comp. Chem*., 2008, in print) and atomistic models of lignin and cellulose were constructed and combined to form an initial model for lignocellulose. Molecular dynamics simulation is performed to examine structural properties of the system. The results of the simulation are compared with SANS data and the computational models are further refined to achieve better agreement with experiments. Development of a simplified modeling approach for the large scale structural organization of lignocellulose has begun. Switch grass, wheat, and other cellulose sources are being cultivated in deuterated media to provide perdeuterated lignocellulose that will enable higher contrast with neutron scattering and examination of surface accessibility.

This work is funded by the U. S. Department of Energy, Office of Biological and Environmental Research Genomics:GTL Program under FWP ERKP704, "Dynamic Visualization of Lignocellulose Degradation by Integration of Neutron Scattering Imaging and Computer Simulation".

GTL

Synchrotron Infrared Spectromicroscopy of Cellulose Degradation Strategies of Living Cellulolytic Bacteria

Hoi-Ying N. Holman[1]* (hyholman@lbl.gov), Zhao Hao,[1] Ping Hu,[1] Robin Miles,[2] Albert Pisano,[3] Inkyu Park,[3] and Sun Choi[3]

[1]Chemical Ecology and Spectromicroscopy Group, Lawrence Berkeley National Laboratory, University of California, Berkeley, Calif.; [2]Microfluidics and Chemical/Bio Devices Group, Lawrence Livermore National Laboratory, Livermore, Calif.; and [3]Micro-Electromechanical Systems Group, University of California, Berkeley, Calif.

Project Goals: Our goal is to provide a new research tool for investigating, directly and in a single experiment, the conversion of biomass from its deconstruction to its conversion to ethanol with unparalleled resolution. To achieve this goal, we have designed technology research and development efforts to meet the following specific aims. 1) to develop an integrated set of non-invasive micro-analytical and imaging technologies to directly characterize lignocellulose degradation by microorganisms in situ; 2) to use these technologies to advance the fundamental understanding of lignocellulose degradation by currently known microorganisms and microbial communities; and 3) to design and confirm the strategy for integrating the SIR technology of live bacterial cells and communities with 3D tracking fluorescence microscopy of single molecules.

We are developing a synchrotron radiation-based Fourier transform infrared (SR-FTIR) spectromicroscopy technology with uninterrupted and non-invasive *in vivo* tracking capabilities. Progress in sequencing and in structural and functional genomics of cellulolytic bacteria must be followed by studying both their strategies of actions on solid substrates such as lignocellulose, inhibitory effects by end products of cellulose degradation, and their adaptation response to toxicity from the fermentation products, such as ethanol. Non-invasive technologies to study such lignocellulose-bacteria systems are not well established.

Our goal is to provide a new research tool for investigating, directly and in a single experiment, the conversion of biomass from its deconstruction to its conversion to ethanol with unparalleled resolution. This technology will have a temporal resolution of seconds-to-minutes and a spatial resolution of several micrometers. Essential to this endeavor is the development of experimental technologies and methods that are robust and allows for the comprehensive measurement and analysis of lignocellulose degradation and the definition of the microbial role. A microfluidic incubation platform and a multi-phasic array of nanowire sensors are being developed to provide a controlled experimental environment that can probe cellular chemistry and manipulate experimental conditions, rather than simply observing processes. Microbial cells and their functions are sensitive to immediate environmental conditions. Genetic strategies are being developed for using model microbes that allow direct confirmation of their viability, metabolic activity, and normal functions inside the microfluidic system during their exposure to SR-FTIR spectromicroscopy measurements.

Today, we are employing lignocellulose-conversion by the mesophilic *Clostridium cellulolyticum* as our model system with which to develop and optimize the selection and tracking of spectral markers, and the performance of synchrotron infrared spectromicroscopy technologies. *C. cellulolyticum* is a representative of anaerobic, mesophilic, cellulolytic bacterium which producing complex cellulose systems (cellulosomes). The cellulosome is thought to allow concerted enzyme activity in close proximity, enabling optimum synergism between the cellulases. An easier DNA transformation system and more biochemical and physiological data make the study of *C. cellulolyticum* valuable. Since the genetic systems are not yet well-established in *Clostridia*, reporter strains are not readily available, and gene/pathway modifications are not easily done. To thoroughly test the integrity and functions of these living cells in the SIR technologies, we are using *E. coli* reporter strains to provide complementary information.

GTL

A New Solution-State NMR Approach to Elucidate Fungal and Enzyme/Mediator Delignification Pathways

Kenneth E. Hammel[1]* (kehammel@wisc.edu), Daniel J. Yelle,[1] John Ralph,[2] Dongsheng Wei,[1] and Alexander N. Kapich[1]

[1]USDA Forest Products Laboratory, Madison, Wis.; and [2]Dept. of Biochemistry, University of Wisconsin, Madison, Wis.

Project Goals: The mechanisms that lignocellulose-degrading fungi employ to bypass lignin and access the polysaccharides in plant biomass are not well understood, in part because the chemical changes they cause in lignin have not been well characterized. We are applying new methods for the complete dissolution of lignocellulose to identify diagnostic degradative reactions in lignin by solution-state NMR spectroscopy.

Brown rot basidiomycetes remove cellulose from wood efficiently, even though this sugar polymer is initially shielded by a biochemically recalcitrant barrier of lignin. During this process, the lignin appears to remain in situ, which raises the question of how the polysaccharide-degrading systems of brown rot fungi circumvent the lignin to access their substrates. New results based on solution-state NMR analysis of ball-milled, dissolved, brown-rotted wood are now available to clarify this picture. We obtained short-range ^{1}H-^{13}C (HSQC) spectra of aspen degraded by the brown-rotter *Postia placenta*, and found that the residual material was about 25% deficient in the major arylglycerol-β-aryl ether structure of lignin, relative to the methoxyl content of the sample. To identify some of the processes responsible for lignin depletion, we next performed additional NMR analyses on decayed, ball-milled aspen that had been treated with a cellulase mixture to remove polysaccharides. A ^{13}C NMR spectrum of this enzyme-treated sample showed that it contained benzoic acid and benzaldehyde residues that were absent in the undecayed aspen. Moreover, a short-range ^{1}H-^{13}C (HSQC) spectrum of the same sample displayed signals characteristic of phenylglycerol residues that were absent in the undecayed wood. The identity of the benzoic acids and benzaldehydes was confirmed in a long-range ^{1}H-^{13}C (HMBC) spectrum, and the identity of the phenylglycerols was confirmed in a total correlation spectroscopy (TOCSY) experiment. The presence of benzoic acid and benzaldehyde residues in the brown-rotted wood shows that its lignin sidechains had been cleaved between C_{α} and C_{β}, whereas the presence of phenylglycerols shows that intermonomer ether linkages between lignin sidechains had been cleaved. Our results demonstrate that *Postia placenta* is actually ligninolytic, contrary to the prevailing view of fungal brown rot. Presumably, ligninolysis is required for fungal polysaccharide hydrolases to operate effectively, but it is interesting, in light of current efforts to improve methods for biomass utilization, that the resulting oxidized polymer does not need to be removed for sugars to be released.

GTL

In Vivo Mapping of ROS Produced by Wood Decay Fungi during Early Colonization

Christopher G. Hunt[1]* (cghunt@fs.fed.us), Kenneth E. Hammel[1,2] (kehammel@wisc.edu), and Don Carl Jones[1]

[1]U.S. Forest Service Forest Products Laboratory; and [2]University of Wisconsin, Madison, Wis.

Project Goals: Develop a modular system to image and quantify ROS (or other metabolite) from fungi in their natural substrate with minimal disturbance. Use the knowledge of ROS concentrations to understand the mechanisms at work during incipient decay.

Background

Wood decay fungi are successful in removing carbohydrate from wood, despite the presence of lignin. We hope to improve prospects for woody biomass saccharification by understanding the mechanisms of fugal wood decay.

To make lignocellulosics permeable to enzymes, filamentous fungi use a variety of small, diffusible reactive oxygen species (ROS) such as hydroxyl radicals, peroxyl radicals, and possibly phenoxy radicals. These radicals diffuse into the cell walls and initiate biodegradative radical reactions. When lignin is the target, radical attack results in various extents of oxidation and depolymerization.

These small diffusible oxidative species are important tools used by filamentous fungi to make the cell wall accessible to enzymes. Despite this, we have a poor knowledge of how these oxidants are spatially distributed in biodegrading lignocellulose relative to the fungal hyphae that produce them. The goal of this project is to remedy this deficit through fluorescence microscopy of newly designed sensors that will serve as in situ reporters of biodegradative radical production. We will use these sensors to produce oxidative maps that will help us to understand how fungi generate ROS and how they use these ROS to make cell walls more accessible to enzymes.

Method

We are attaching fluorescent dyes to silica beads. Our first bead has BODIPY 581/591® on a 3µm porous HPLC bead. This dye's emission changes irreversibly from red to green upon oxidation by ROS. The ratio of red to green emission provides a quantitative measure of the cumulative oxidation at that point in space. Dyes with reactivity to specific ROS, pH, or other metabolites of interest are envisioned.

There are many advantages gained by fixing the dye to bead. We design the bead to emit two fluorescent signals, so that the ratio of the two signal intensities provides quantitative information. Immobilized dyes are prevented from moving

after reaction, so partitioning is impossible, they cannot be ingested, and the fluorescence from the dye is clearly distinguishable from background.

Beads are placed on wood samples and imaged with a confocal microscope during fungal colonization. Images can be analyzed to provide the analyte concentration maps as well as an overlay of the location of fungal hyphae.

Results
We see large changes in the fluorescence of our beads with fungal oxidation, as shown in 1. Ideally, we would be able to see the ROS gradient around a single hypha. We developed a silica gel culture system so that the fungus could not metabolize the water agar gel and multiply too quickly. We have succeeded in observing an ROS gradient around individual hypha, as shown in Figure 2. Individual beads show oxidation through a lower average ratio of red to green emission, and by displaying a pattern of oxidized pixels on the perimeter while the core of the bead has higher red to green ratios.

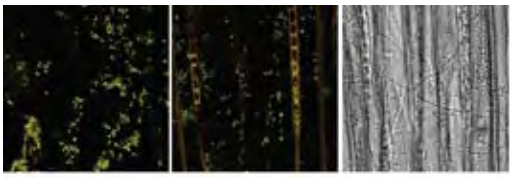


Figure 1. Fluorescent beads no fungus (left) and 3 days after inoculation with *Phanerochaete chrysosporium* (center). Note change in bead color with oxidation. Transmission image shows fungal hyphae (Right) Wood cellular structure is visible via autofluorescence (center).

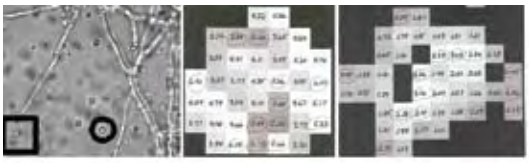


Figure 2. Transmission image (left) of fungal hyphae growing over gel adjacent to wood. Oxidized bead is circled and unoxidized bead is squared. Unoxidized bead (center photo) has bright and dark pixels, representing degree of oxidation, randomly distributed. Oxidized bead (right) is most oxidized (darkest pixels) around perimeter while less oxidized light pixels are all in center. Note the average pixel value is also lower.

Summary
We are able to measure oxidation produced by colonizing hyphae at a micron scale. We are currently generating micron scale oxidation maps to determine the ROS gradient as a function of distance. Next we plan to compare the oxidative profile of brown rot, white rot, and nondecay fungi.

GTL

Stimulated Raman Scattering as an Imaging Tool for Lignocellulosic Biomass Conversion

Brian G. Saar, Yining Zeng, Christian W. Freudiger, Marcel G. Friedrich, Gary R. Holtom, Shi-you Ding, and **X. Sunney Xie*** (xie@chemistry.harvard.edu)

Harvard University, Cambridge, Mass.

Project Goals: To improve the conversion efficiency of ligno-cellulosic biomass to biofuels with SRL microscopy.

Targeted lignin modification in bioenergy crops could improve conversion efficiency of lignocellulosic biomass to biofuels. In order to understand the practical consequences of genetic mutations of different crops and variations in the degradation procedures, real-time imaging of the two key chemical components involved in the process, lignin and cellulose, is needed. We demonstrate the use of simultaneous two color stimulated Raman loss (SRL) imaging to address this need. The SRL technique is capable of real-time chemical imaging of multiple components in plant cell wall samples without the use of labels. SRL imaging is background free and linearly dependent on the chemical concentration, allowing for straightforward and rapid quantification of both the lignin and cellulose in the plant samples with sub-micron spatial resolution and high time resolution. We show applications to a variety of plant samples that are candidates for biofuels, including engineered mutants of alfalfa and corn stover with varying lignin content. The superior sensitivity and image contrast in SRL microscopy make it an ideal tool for studying lignocellulosic degradation.

GTL

Probing the Architecture of the Plant Cell Wall during Deconstruction in Single Cells from *Zinnia elegans*

C.I. Lacayo[1]* (lacayo2@llnl.gov), H.-Y. Holman,[2] M. Hwang,[1] A. Hiddessen,[1] A.J. Malkin,[1] and **M.P. Thelen**[1] (mthelen@llnl.gov)

[1]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.; and [2]Ecology Dept., Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: To use a combination of imaging approaches, including atomic force microscopy (AFM), synchrotron radiation Fourier-transform infrared spectroscopy (SR-FTIR), and fluorescence microscopy, to investigate the structural and dynamic features of plant cell wall biogenesis and degradation.

The plant cell wall is primarily composed of lignocellulose, a non-food source of fermentable sugars, which can become

the starting material for the production of a variety of biofuels. The cell wall is also a highly complex and dynamic structure, which is not amenable to efficient commercial degradation using currently available technologies. Our goal in this project is to understand how the cell wall changes in response to chemical, enzymatic and microbial treatments, using a combination of imaging platforms, including atomic force microscopy (AFM), synchrotron radiation Fourier-transform infrared spectroscopy (SR-FTIR), and fluorescence microscopy. These techniques provide the capability to image changes in the organization of plant cell surfaces and reveal corresponding chemical composition, thus enabling a more comprehensive understanding of cell wall architecture.

To relate our imaging studies to bioenergy production from plant biomass with high lignocellulose content, we use a *Zinnia elegans* model culture system in which single cells develop lignocellulose-rich walls. *Zinnia* leaf mesophyll cells are induced to gradually differentiate in semi-synchrony into single tracheary elements (TEs), individual components of xylem tissue (panel A in the figure). Mature TEs possess large secondary cell wall thickenings, which are rich in lignocellulose and represent the bulk of woody biomass. These secondary cell wall thickenings become deposited below the primary cell wall and are highly ordered, creating helical or reticulate patterns (panel A). Cultured *Zinnia* TEs represent an advantage over plant tissues by providing a more homogeneous source of lignocellulosic material and by enabling the examination of both primary and secondary cell wall deconstruction in individual cells. In addition, the *Zinnia* cell culture system permits us to probe the cell wall at various stages of the TE differentiation process.

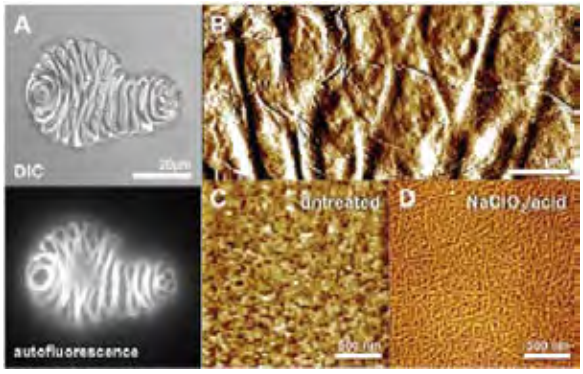


Figure. (A) Differential interference contrast (DIC) and autofluorescence images of a differentiated tracheary element from *Zinnia elegans* show a reticulate pattern of the secondary wall thickenings. Secondary cell wall autofluorescence is attributed to lignin. (B) A low-resolution AFM (amplitude) image of a tracheary element also reveals the presence of wall thickenings. (C) A high-resolution AFM (contrast-enhanced height) image shows granular structures on the surface of a cultured untreated tracheary element. (D) Cellulose fibrils become apparent after a chemical treatment.

We have observed that surface organization and composition of *Zinnia* TEs change dramatically after various chemical treatments. As illustrated in panel C in the figure, the cell walls of untreated TEs were found by AFM to have pronounced granular structures, ranging in size from about 20 to 50 nm. Chemical treatments resulted in the removal of this outermost granular material, revealing the underlying network of cellulose fibrils, which ranged in diameter from approximately 10 to 20 nm (panel D). Distinctive changes in TE composition following chemical treatments were also examined by SR-FTIR.

We are currently developing experimental approaches for high-resolution topographical and structural characterization of the cell wall in TEs and their dynamic response to chemical, enzymatic and microbial treatments. We expect that these approaches will enable us to develop molecular-scale models of the architecture and mechanisms of degradation of the plant cell wall.

This work was funded by the U.S. Department of Energy GTL Program project number SCW0738 and performed under the auspices of the DOE by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

GTL

Integrated Nondestructive Spatial and Chemical Analysis of Lignocellulosic Materials during Pretreatment and Bioconversion to Ethanol

Gary Peter* (gfpeter@ufl.edu), Choong Heon Lee, Kyle Lundsford, Steve Blackband, Lonnie Ingram, and Richard Yost

University of Florida, Gainesville, Fla.

Project Goals: The objectives of this proposed work are to 1) use advanced high resolution magnetic resonance microscopy (MRM), micro and nano x-ray computed tomography (x-ray CT), electron microscopy and imaging mass spectrometry (IMS) to quantify changes in the architecture, porosity, permeability, surface area, pore size, interconnectivity, chemical organization and composition of corn stover and particularly *Populus* and pine wood chips during pretreatment and enzymatic degradation, and; 2) integrate the results from these different quantitative imaging methods into a model for disassembly of the plant cell wall during pretreatment and bioconversion.

Our *long-term goal* is to develop a quantitative structural model for changes that occur in the organization and chemical composition of plant biomass during pretreatment, enzymatic degradation and bioconversion to ethanol or other products. The *objectives* of this proposed work are to 1) use advanced high resolution magnetic resonance microscopy (MRM), micro and nano x-ray computed tomography (x-ray CT), electron microscopy and imaging mass spectrometry (IMS) to quantify changes in the architecture, porosity, permeability, surface area, pore size, interconnectivity, chemical organization and composition of bagasse and particularly *Populus* and pine wood chips during pretreatment and enzymatic degradation, and 2) integrate the results

* Presenting author

from these different quantitative imaging methods into a model for disassembly of the plant cell wall during pretreatment and bioconversion. We have standardized the methods for imaging biomass with MRM, x-ray CT and IMS.

IMS: Standardized method for sample preparation and mounting using cryosectioning has been developed for biomass. A number of matrices have been tested and work well with sections of *Populus* wood. Pretreatment of wood with dilute sulfuric acid leads to very complex MS spectra, creating challenges for peak identification. Pure cellulose, hemicelluloses and lignin are being tested to identify characteristic oligomer peaks to focus searches.

MRM: Excellent image quality is obtained from *Populus* wood and bagasse samples using T2 and diffusion weighted modes. In T2 images, vascular bundles appear dark consistent with the knowledge that lignified cells contain limited free water. A chemical shift has been found in some samples and the cause for this is under investigation.

X-ray CT: Excellent images have been obtained at high resolution from *Populus*, pine, and bagasse samples. In addition to the basic density, images are readily segmented and the material and airspace sizes can be quantified.

Current work aims to validate quantitative analysis of each imaging method with the same biomass samples.

GTL

Single-Molecule Studies of Cellulose Degradation by Celluosomes

Haw Yang* (hyang@lbl.gov), Jamie Cate (jcate@lbl.gov), Chang-Li Chen, Padma Gunda, Dan Montiel, and Hu Cang

Lawrence Berkeley National Laboratory and University of California, Berkeley, Calif.

Project Goals: Ethanol production from biomass is one of the few promising solutions to our current energy problem. As a renewable energy source, this approach is hindered by the rate-limiting step of lignocellulose degradation. In nature, this step is accomplished enzymatically by the extracellular cellulosomes from anaerobic microbes. The mechanism, however, has been nearly impossible to study at the biophysical level due to the multiple-scale nature of the problem. The single-molecule approach affords the direct observation of molecular processes within a heterogeneous distribution to shed light on this important issue. Yet, technologies to image individual molecular machines in temporally and spatially heterogeneous setting are not well established.

In order to probe function and dynamics of molecular machines in situ, we propose to develop a single-molecule spectrometer with the capability to track individual fluorescent particles as they move in three dimensions (3D). This spectrometer will provide unprecedented time and spatial resolution for imaging single molecules in a complex environment. We will develop and optimize biomolecule tagging, single-molecule assays, and 3D single-particle tracking technologies. These technologies will help establish single-molecule spectroscopy with 3D positioning as a general approach to study molecular machines in complex environments.

A cellulosome is a large extracellular supramolecular complex that is produced by anaerobic microbes to enzymatically decompose crystalline cellulosic polymers and plant cell walls. It consists of a scaffolding protein that accommodates other essential protein and enzyme components for cellulose degradation. They include carbonhydrate-binding modules for attachment to the solid cellulose substrate, various glycoside hydrolases to efficiently hydrolyze a heterogeneous substrate, and in some cases, anchoring proteins to attach cellulosomes to the bacterial cell surface. The mechanism of its function is poorly understood due to the complexity of the cellulosome itself and the natural environment in which it functions. Outstanding issues include the location and the manner in which the cellulosome is assembled, the distribution in the cellulosomal composition, and the dynamic interactions between cellulases and the insoluble substrate, to name a few.

These questions are very difficult to address quantitatively using conventional, ensemble-based methods due to the multiple layers of complexity involved. The convoluted spatio-temporal dynamics in cell and cellulosome interaction with insoluble substrates make it very hard to quantitatively study the various molecular dynamics of a functioning cellulosome. We anticipate that the single-molecule approach, due to its capability of directly monitoring the individual processes from a distribution, will prove invaluable in efforts to unravel how microscopic, molecular interactions impact macroscopic behavior in plant cell wall degradation.

In order to study the various processes involved in lignocellulosic degradation by cellulosomes, we are developing a single-molecule spectrometer with the capability to track individual fluorescent particles in three dimensions. We have developed three-dimensional single-particle tracking spectroscopy instrument that allows us to follow the 3D movements of individual luminescent nanoprobes and at the same time perform spectroscopic investigation. The capabilities will be briefly outlined. We have also developed protocols for imaging individual cellulosomes acting on well prepared crystalline cellulose fibers. Preliminary data will be presented. The mechanistic insights are expected to have a direct impact on the improvement and engineering of tailored biomass depolymerization systems.

Systems Biology and Metabolic Engineering Approaches for Biological Hydrogen Production

GTL

Metabolic Modeling for Maximizing Photobiological H_2 Production in Cyanobacteria

Alexander S. Beliaev[1]* (alex.beliaev@pnl.gov), Grigoriy E. Pinchuk,[1] Oleg V. Geydebrekht,[1] Michael H. Huesemann,[2] Elizabeth H. Burrows,[3] Trang Vu,[4] Jennifer L. Reed,[4] Andrei L. Osterman,[5] John R. Benemann,[6] and Jim K. Fredrickson[1]

[1]Pacific Northwest National Laboratory, Richland, Wash.; [2]Pacific Northwest National Laboratory Marine Science Laboratory, Sequim, Wash.; [3]Oregon State University, Corvallis, Ore.; [4]University of Wisconsin, Madison, Wis.; [5]Burnham Institute for Medical Research, La Jolla, Calif.; and [6]Benemann Associates, Walnut Creek, Calif.

Project Goals: In this project, we seek to improve our understanding of H_2 production by a diazotrophic unicellular cyanobacterium *Cyanothece sp.* strain ATCCC 51142 using a metabolic modeling approach for simulating the fundamental metabolism of indirect biophotolysis, as well as identifying the main metabolic and regulatory controls in this organism. From this, the potential for H_2 production by indirect biophotolysis in this organism will be assessed based on imposing new constraints to redirect the low redox potential electron transport pathways from normal metabolism towards H_2 production from accumulated carbohydrates. As a result, it will provide an in silico tool for manipulating such microorganisms to act as catalysts for solar energy conversion to H_2 and potentially allow for a development of a highly efficient H_2 production process.

Advances in microbial genome sequencing and functional genomics are greatly improving the ability to construct accurate systems-level models of microbial metabolism and to use such models for metabolic engineering. With the increasing concerns over the reliance on fossil fuels, there is a revitalized interest in using biological systems for producing renewable fuels. Genomics and metabolic engineering hold great promise for the rational design and manipulation of biological systems to make such systems efficient and economically attractive.

In this project, we are developing a constraint-based metabolic model for marine diazotrophic cyanobacterium *Cyanothece* ATCC 51142. The initial set of reactions, used for the first draft of our metabolic model, was deduced from genomic annotations generated using the RAST (Rapid Annotation using Subsystems Technology) server. A draft reconstruction was generated in SimPheny® by Genomatica (San Diego, CA) based on sequence comparisons between *Cyanothece* and organisms with genome-scale metabolic models. This reconstruction was a starting point, upon which organism specific pathways were added, and gene-protein-reaction associations were closely evaluated. The current model includes 569 metabolic and transport genes, 542 proteins, and 715 reactions. The following pathways are accounted for in the current model: central metabolism, fatty acids biosynthesis, amino acid synthesis, nucleotide synthesis, cofactor biosynthesis, oxidative phosphorylation, and photosynthesis. The number of metabolic gaps (metabolites which can only be produced or only consumed) has been reduced from 484 in the original draft reconstruction to 302 and is expected to decrease further as other pathways (such as nitrogen fixation and carbon fixation) are being incorporated into the current model. In addition, the biomass formation reaction is being constructed based on experimental measurements and the reaction will help significantly reduce the number of gaps in the model.

Conditions promoting H_2 production by *Cyanothece* ATCC 51142 are also being studied in order to develop a strategy for maximizing H_2 production using metabolic modeling approach. Preliminary results showed that this cyanobacterial strain is not capable of generating significant amounts of H_2 from stored glycogen under dark fermentative conditions. However, illumination of *Cyanothece* ATCC 51142 cultures exposed to Ar atmosphere and deprived of nitrogen source significantly increased the rate of light-driven H_2 production. No O_2 was produced in course of more than 60 hours hydrogen photoproduction. Interestingly, the addition of DCMU, a photosystem II inhibitor, did not affect this process, whereas presence of N_2 in atmosphere or ammonia in the medium effectively prevented H_2 appearance. Further biochemical analysis revealed that nitrogenase was the enzyme primarily responsible for light-driven H_2 production by *Cyanothece* ATCC 51142 under the experimental conditions. Current hypothesis of metabolic pathways involved in nitrogenase mediated H_2 production includes generation of ATP by cyclic electron transport in photosystem I, whereas stored carbohydrates are used as source of electrons. The predicted pathways and experimental data will be incorporated into the metabolic model of *Cyanothece* ATCC 51142 to identify the means for maximization of H_2 production by this organism.

* Presenting author

GTL

A High-Throughput Genetic Screen for Large Subunit Hydrogenase (*hoxH*) Enzyme Presence in Cyanobacteria Using Whole-Cell Duplex PCR and Novel Amplification Parameters

Nicholas J. Skizim (nskizim@princeton.edu), Derrick R.J. Kolling,* Diana Chien, and **G. Charles Dismukes**

Dept. of Chemistry, Princeton University, Princeton, N.J.

Project Goals: The goal of this project is the development of a high-throughput screening method for the enzymatic presence of hydrogenase in cyanobacteria.

In designing a screen for hydrogen production capabilities in cyanobacteria, it is advantageous to use a method which has high-throughput as well as high confidence in identifying positive hits. Here we describe degenerate primers amplifying a portion of the large subunit of the hydrogenase complex (*hoxH*) designed from protein motifs which bind the [Ni-Fe] active site in the complex and are conserved throughout cyanobacteria. These oligonucleotides contain flanking restriction sites to increase target specificity as well as to facilitate subsequent cloning/sequencing. The screen relies on Duplex PCR to concurrently amplify fragments of both *psbA* (the photosystem II D1 protein) along with *hoxH* to provide a positive control on the reaction conditions, enzyme activity, and whole-cell template. Touchdown PCR is in use to achieve the highest level of specificity from the degenerate primers and the PCR cocktail is tailored as well to (a) increase specificity, and (b) be compatible with whole-cell template methods. A preliminary screen of five cyanobacterial strains (*Cyanothece Miami BG043511, Synechocystis PCC6803, Synechococcus PCC7002, Arthrospira maxima, and Gloeobacter violaceus)* has verified the presence and absence of hydrogenase genes as expected, illustrating proof of method. Data is also presented showing the effect of various cellular pre-treatments on the amplification. Further, the possibilities of a fluorescent based detection instead of agarose gel electrophoresis and ethidium bromide stain are discussed.

This work has been supported by the Department of Energy – Genomics:GTL, DE-FG02-07ER64488, as well as the Air Force Office of Scientific Research, AFOSR-MURI- FA9550-05-1-0365.

GTL

Probing the Interdependence of Hydrogenase- and Nitrogenase-Dependent Hydrogen Production in *Cyanothece* sp. Miami BG043511 by Genetic Deletion and Mutant Analysis

Nicholas J. Skizim[1]* (nskizim@princeton.edu), Randall Kerstetter,[2] Todd Michael,[2] and **G. Charles Dismukes[1]**

[1]Dept. of Chemistry, Princeton University, Princeton, N.J.; and [2]Waksman Institute of Microbiology, Rutgers University, Piscataway, N.J.

Project Goals: The goal of this project is for the development of quantitative tools to measure hydrogen production in photosynthetic bacteria.

The unicellular diazotrophic cyanobacterium *Cyanothece* sp. Miami BG043511 is unique in being the sole reported cyanobacterium possessing nitrogenase and bidirectional hydrogenase but lacking an "uptake" hydrogenase. In addition to the bidirectional enzyme, the nitrogenase can produce hydrogen. Because of the complex anaerobic energy balance in this organism, including the ATP demand of nitrogen fixation, it is difficult to chemically inhibit one of the two hydrogen-producing enzymes without indirectly affecting the other. Therefore, we aim to create mutants deficient in each of the above named enzymes, including a double mutant in which both are deleted. Initial work relying on degenerate primers and inverse PCR has given us the complete sequence of the nitrogenase iron protein (*nifH*) as well as a partial sequence homologous to *hupL*, the large subunit of the uptake hydrogenase. We have collected sequence data from this organism using Applied Biosystem's SOLiD sequencing techniques and are in the process of assembling the entire genome. A rough assembly of the genome along with Sanger sequencing to fill in the gaps has confirmed this presence of a complete *hupL*. Because this strain is claimed to lack activity of the uptake hydrogenase and this observation is in contradiction to that claim, we are in the process of assessing the claim of zero uptake hydrogenase function. A mutant (Δ*hupL*) is being constructed to determine whether the genetic presence of this enzyme influences the phenotype in any way. Hydrogen production from this strain has been measured on an ultra-sensitive clark-type electrode revealing multiple kinetic phases of hydrogen production under fermentative conditions. This hydrogen production, as well as the excretion of other fermentative end-products, has been investigated both when the culture is concurrently fixing nitrogen as well when it is not. The data from this shall be presented.

This work has been supported by the Department of Energy – Genomics:GTL, DE-FG02-07ER64488, as well as the Air Force Office of Scientific Research, AFOSR-MURI- FA9550-05-1-0365.

Microalgal Biomass as Feedstock for Production of Hydrogen and Methane Fuels

Mark B. Smith[1]* (mbsmith@princeton.edu), Damian Carrieri[1] (carrieri@princeton.edu), Jun Cheng[2] (juncheng@zju.edu.cn), Gennady Ananyev[1] (anayev@princeton.edu), and **G. Charles Dismukes**[1] (dismukes@princeton.edu)

[1]Princeton University, Princeton, N.J.; and [2]Zhejiang University, Zhejiang Province, China

http://www.princeton.edu/~biosolar/

As an alternative to land-based crops for biofuels, Aquatic Microbial Oxygenic Phototrophs (AMOPs), such as algae, cyanobacteria and diatoms have several potential advantages. AMOPs are inherently more efficient solar collectors, use less or no land, and can be converted to usable fuels using simpler technologies than cellulose (1). One developed strategy for conversion of biofeedstocks to methane and hydrogen is anaerobic digestion using microbes from sewage sludge. This has been previously demonstrated for glucose and starch containing feedstocks from land-based crops (2-3). Here we illustrate the use of AMOPs as an alternative feedstock for such processes. AMOPs are buoyant cells that are completely devoid of the recalcitrant biopolymers that comprise terrestrial crops (cellulose, lignin, hemi-cellulose) and enriched in energy precursors. *Arthrospira maxima* is a filamentous cyanobacterium that grows to high biomass densities in open aquifers and is permissive of high salinity, high pH and high carbonate concentrations, conditions that suppress microbial contamination. *Arthrospira* species have been grown in large outdoor facilities for commercial purposes and demonstrated to grow on fertilized sea water (4). The high conversion efficiency of cyanobacterial biomass to methane and hydrogen using our process will be discussed.

References

1. G.C. Dismukes, D. Carrieri, N Bennette, G. M. Ananyev, and M.C. Posewitz. (2008) "Aquatic phototrophs: efficient alternatives to land-based crops for biofuels" *Current Opinion in Biotechnology* 19: 235-240
2. Binfei Xie, Jun Cheng, Junhu Zhou, et al. Cogeneration of hydrogen and methane from glucose to improve the energy conversion efficiency. *International Journal of Hydrogen Energy*, 2008, 33(19): 5006-5011.
3. Binfei Xie, Jun Cheng, Junhu Zhou, et al. Production of hydrogen and methane from potatoes by two-phase anaerobic fermentation. *Bioresource Technology,* 2008, 99(13): 5942-5946.
4. M. R. Tredici, T. Papuzzo, and L. Tomaselli (1986) "Outdoor mass culture of *Spirulina maxima* in sea-water" *Applied and Microbiology and Biotechnology* 24: 47-50

Supported by AFOSR-MURI- FA9550-05-1-0365 and Princeton University.

Development of Biologically-Based Assays to Study Rate-Limiting Factors in Algal Hydrogen Photoproduction

Alexandra Dubini,[1] Matthew Wecker,[2] and **Maria L. Ghirardi**[1]* (maria_ghirardi@nrel.gov)

[1]National Renewable Energy Laboratory, Golden, Colo.; [2]GeneBiologics, LLC, Boulder, Colorado

Photobiological H_2 production from water is a clean, non-polluting and renewable technology. Although the potential light conversion efficiency to H_2 by biological organisms is theoretically high (about 10%), the system is currently limited by biochemical and engineering constraints. These limitations include (but are not restricted to) the extreme O_2 sensitivity of the biological hydrogenases and the low availability of reductants to the hydrogenase due to the existence of competing metabolic pathways.

Our research addresses the O_2 sensitivity issue by developing a new, biologically-based assay to screen large microbial populations for improved H_2-production properties. This novel assay is based on the H_2-sensing properties of systems found in nitrogenase-containing photosynthetic bacteria. We will validate the new assay by using it to screen mutants generated through directed-evolution techniques for O_2 tolerant [FeFe]-hydrogenases. The hydrogen-sensing system in *R. capsulatus* is being optimized as an assay of heterologous hydrogen production by the bacteria and will then be used to screen mutants generated through directed-evolution techniques for O_2 tolerant [FeFe]-hydrogenases. The hydrogenases of *Clostridium acetobutylicum* and *Bacteroides thetaiotaomicron*, along with their respective assembly proteins have been introduced into broad host range vectors and are being shuttled into the photosynthetic bacteria *Rhodobacter capsulatus*.

To address the issue of competitive metabolic pathways with H_2 production, we will adapt the yeast two-hybrid assay to measure the interactions between different ferredoxin isoforms present in *Chlamydomonas reinhardtii* with different proteins known to accept electrons from ferredoxin in most photosynthetic and fermentative organisms. We started examining the direct interaction of the HYDA2 hydrogenase and other partners using a yeast two-hybrid system. The cDNA library has been prepared and a first screen has been done. Some prey clones have been picked and are being tested for interaction strength by growth tests and β-galactosidase assay.

This project will develop techniques that will drive a deeper understanding of algal H_2 metabolism and accelerate the development of future photobiological H_2-production catalysts and organisms.

GTL

Novel Hydrogen Production Systems Operative at Thermodynamic Extremes

Michael J. McInerney,[1] Jesscia Seiber,[1] Ralph S. Tanner,[1] and **Robert P. Gunsalus**[2]* (robg@microbio.ucla.edu)

[1]University of Oklahoma, Norman, Okla.; and [2]University of California, Los Angeles, Calif.

http://www.mimg.ucla.edu/faculty/gunsalus/

Project Goals: The goals of our collaborative project are to develop new research strategies, model organisms, and research resources addressing the needs of the Genomics:GTL program in the area of biohydrogen production. We will apply a suite of molecular, bioinformatics, and biochemical tools to: i) interrogate the thermodynamically limiting steps of H_2 production from fatty and aromatic acids in syntrophic communities and in pure cultures, ii) develop a new microbial model system that generates high H_2 concentrations (over 17% of the gas phase) with high hydrogen yields (3+ mol H_2 per mole glucose), and iii), perform systems-based studies of biohydrogen production in model anaerobe consortia to identify key metabolic steps. The results of these studies will greatly expand our ability to predict and model systems for H_2 production in novel anaerobes that are currently very poorly understood.

The goals of this collaborative project are to develop new research strategies to address the Genomics:GTL program needs in the area of bio-hydrogen production. This includes the delineation of the molecular machinery involved in hydrogen production from thermodynamically difficult substrates, as well as the characterization of new microbial model systems that generate high H_2 concentrations approaching 17% of the gas phase.

To identify the molecular machinery for hydrogen production in model syntrophic microorganisms we are performing genomic analysis with *Syntrophus wolfei.* This bacterium is representative of an important but poorly understood class of hydrogen producing organisms that are capable of syntrophic fatty and aromatic acid metabolism when co-cultured with suitable microbial partner(s). Their ability to produce H_2 requires reverse electron transport where energy input is needed.

Analysis of the *S. wolfei* genome predicts that it encodes multiple hydrogenase enzymes, two of which are predicted to be cytoplasmic hydrogenases, and one is predicted to an externally located hydrogenase. One of the soluble Fe-only hydrogenase genes is clustered with genes for NADH dehydrogenase. This arrangement is also found in two other sequenced organisms capable of syntrophy, *Pelotomaculum thermopropionicum* and *Syntrophobacter fumaroxidans.* Thus, one mechanism for hydrogen production during syntrophy maybe to form a complex that directly couples NADH oxidation with hydrogen production. The externally-oriented is predicted to be membrane-associated, ferredoxin hydrogenase (E.C. 1.12.7.2). One of ORFs in this cluster was most similar to a gene for cytochrome b_{556}, suggesting that this hydrogenase could produce hydrogen from a reduced quinone intermediate by reverse electron transfer. Cells of *Syntrophomonas wolfei*, each grown in coculture with *Methanospirillum hungatei,* were separated from the methanogenic partner by Percoll density centrifugation and the amounts of externally-oriented hydrogenase and total hydrogenase activities were measured. Most of the hydrogenase activity was externally-oriented. These data implicate the involvement of membrane-associated hydrogenases when thermodynamically difficult substrates such as fatty and aromatic acids are the substrates.

In another project we are characterizing a newly isolated bacterium called *Anaerobaculum* strain OS1 that can generate H_2 at concentrations up to 17%. This ability suggests the presence of a novel system for H_2 production from carbohydrates not involving NADH where we hypothesize that glucose metabolism involves ferredoxin-linked rather than NADH-linked dehydrogenases and membrane-bound hydrogenase systems. Initial microbiological characterization of strain OS1 reveals phenotypically distinguishable traits from *A. thermoterrenum* and *A. mobile,* and it is designated as a new species: "*Anaerobaculum hydrogenoformans*". In pure culture, strain OS1 produced 3.85 moles of H_2 per mole of glucose and is a much better H_2 producer than either of the two described *Anaerobaculum* species. OS1 also grew syntrophically in the presence of a H_2-consuming methanogen, suggesting the ability for a more complex alternative lifestyle.

GTL

Mechanism of Post-Translational Control of Nitrogenase is Revealed by Analysis of Constitutive Hydrogen-Producing Mutants of *Rhodopseudomonas palustris*

Erin K. Heiniger* (erinkh@u.washington.edu), Yasuhiro Oda, and **Caroline S. Harwood**

Dept. of Microbiology, University of Washington, Seattle, Wash.

Project Goals: The overall goal of this project is to develop and apply techniques in metabolic engineering to improve the biocatalytic potential of the bacterium *Rhodopseudomonas palustris* for nitrogenase-catalyzed hydrogen production. *R. palustris*, is an ideal platform to develop as a biocatalyst for hydrogen gas production because it is an extremely versatile microbe that produces copious amounts of hydrogen by drawing on abundant natural resources of sunlight and biomass. Anoxygenic photosynthetic bacteria, such as *R. palustris*, generate hydrogen and ammonia during nitrogen fixation. The applied use of nitrogenase for hydrogen production is attractive because hydrogen is an obligatory product of

this enzyme and is formed as the only product when nitrogen gas is not supplied. Our challenge is to understand the systems biology of *R. palustris* sufficiently well to be able to engineer cells to produce hydrogen continuously, as fast as possible and with as high a conversion efficiency as possible of light and electron donating substrates. This abstract increases our understanding of how nitrogenase and hydrogen production are regulated in *R. palustris*.

Rhodopseudomonas palustris is a phototrophic proteobacterium that produces hydrogen gas via nitrogenase, the enzyme responsible for biological nitrogen fixation. Nitrogenase expression and activity are tightly controlled by the bacterium because nitrogen fixation is a costly reaction. Previously we described mutants that produce hydrogen constitutively, even when grown in the presence of ammonium, due to activating NifA* mutations in the regulator NifA (1). Ammonium is a biologically available nitrogen source that normally prevents NifA from activating nitrogenase gene transcription. In many bacteria nitrogenase activity is also tightly controlled by DRAT and DRAG enzymes. DRAT inactivates nitrogenase by ADP-ribosylation in response to ammonia, while DRAG removes the modification. NifA does not control the expression of these enzymes in *R. palustris*. We constructed a DRAT mutant and confirmed that DRAT inactivates nitrogenase in *R. palustris*. We also confirmed that DRAT does not switch off nitrogenase activity or hydrogen production in our NifA* mutants in response to ammonia addition. To explain this surprising finding we hypothesized that the hydrogen-producing mutants do not express a small regulatory PII protein that is required to activate DRAT activity in response to ammonia. We constructed mutations in the *glnB*, *glnK1* and *glnK2* genes that encode the PII proteins of *R. palustris*, and determined that GlnK2 is required for the activation of DRAT upon ammonium addition to cultures. In complementary microarray studies, we found that genes controlled by the *ntrBC* regulatory system are not expressed in the hydrogen-producing mutants grown in the presence of ammonium. Among these is the *glnK2* gene. Thus, in addition to constitutively activating nitrogenase gene expression, NifA* mutations also indirectly allow *R. palustris* cells to escape posttranslational inactivation of nitrogenase and produce hydrogen even when grown in the presence of ammonium.

Reference

1. Rey, F. E., E. K. Heiniger and C. S. Harwood. 2007 Redirection of metabolism for biological hydrogen production. *Appl. Environ. Microbiol.* **73**:1665-1671

GTL

Systems Biology of Hydrogen Regulation in *Methanococcus maripaludis*

Erik L. Hendrickson,[1] Qiangwei Xia,[1] William B. Whitman,[2] Murray Hackett,[1] **John A. Leigh**[3]* (leighj@u.washington.edu), and Nitin Baliga[4]

[1]Dept. of Chemical Engineering, University of Washington, Seattle, Wash.; [2]Dept. of Microbiology, University of Georgia, Athens, Ga.; [3]Dept. of Microbiology, University of Washington, Seattle, Wash.; and [4]Institute for Systems Biology, Seattle, Wash.

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

Background

We are engaged in a long-term effort to understand regulatory networks in hydrogenotrophic methanogens, members of the Archaea whose energy metabolism specializes in the use of H_2 to reduce CO_2 to methane. (Many hydrogenotrophic methanogens can use formate as an alternative to H_2 and CO_2). Our studies focus on *Methanococcus maripaludis*, a model species with good laboratory growth characteristics, facile genetic tools, and a tractable genome of 1722 annotated ORFs. Much of our work to date has focused on the response that occurs when supplies of essential nutrients are decreased to growth-limiting levels. Thus, we have studied the responses to H_2 limitation, nitrogen limitation, phosphate limitation, and leucine limitation (using a leucine auxotroph) [1, 2].

Continuous culture of *M. maripaludis*

A key aspect of our approach is the use of continuous culture for maintaining defined nutrient conditions [3]. Chemostats are operated at a constant dilution rate, hence growth rate is constant. Cell density is held constant as well, determined by the supply of the growth-limiting nutrient.

Effect of H_2-limitation on the transcriptome

The effect of limitation by a given nutrient is determined by comparison with at least two other nutrient limitations. Thus, samples from H_2-limited cultures were compared with samples from phosphate-limited cultures and samples from leucine-limited cultures. Array analysis revealed that the mRNA levels most highly affected by H_2 were for those encoding certain enzymes in the methanogenic pathway. Notably, these enzymes use coenzyme F_{420} as electron donor or receptor, highlighting the importance of electron flow through F_{420}.

Effect of H_2-limitation on the proteome

Proteomics was conducted by 2-D capillary HPLC coupled with tandem mass spectrometry. Samples from H_2-limited cultures were compared with samples from phosphate-

limited cultures and samples from nitrogen-limited cultures. Eight percent of the proteome changed significantly with H_2 limitation. Many of the proteins involved in methanogenesis increased. One protein involved in methanogenesis decreased: a low-affinity [Fe] hydrogenase, which may dominate over a higher-affinity mechanism when H_2 is abundant.

Utility of high-coverage quantitative proteomics in *M. maripaludis*

With H_2 limitation, a wider variety of proteins involved in methanogenesis increased than was previously indicated for mRNAs. While a variety of factors are possible, it is likely that part of the explanation lies in the superior ability of the proteomics to discern statistically significant differences. Exhaustive sampling of the proteome was key. When sampling the proteome at or near saturation in terms of qualitative identifications from proteolytic digests, the statistical power to detect a defined fold change between two sets of growth conditions, at a defined level of significance, is quite high for many *M. maripaludis* proteins. How many of these statistically significant abundance changes are of biological relevance is a question we are actively investigating.

Future directions

We plan an extensive study of transcriptomic and proteomic correlates to changes in H_2 metabolism (utilization and production), electron flow, nitrogen fixation, and carbon assimilation, leading ultimately to modeling of regulatory networks. We will also investigate the mechanism of H_2 sensing and regulation.

References

1. Hendrickson EL, Haydock AK, Moore BC, Whitman WB, Leigh JA: Functionally distinct genes regulated by hydrogen limitation and growth rate in methanogenic Archaea. *Proc Natl Acad Sci U S A* 2007, 104(21):8930-8934.
2. Hendrickson EL, Liu Y, Rosas-Sandoval G, Porat I, Söll D, Whitman WB, Leigh JA: Global responses of *Methanococcus maripaludis* to specific nutrient limitations and growth rate. *J Bacteriol* 2008, 190(6):2198-2205.
3. Haydock AK, Porat I, Whitman WB, Leigh JA: Continuous culture of *Methanococcus maripaludis* under defined nutrient conditions. *FEMS Microbiol Lett* 2004, 238(1):85-91.

GTL

Systems Level Approaches to Understanding and Manipulating Heterocyst Differentiation in *Nostoc punctiforme*: Sites of Hydrogenase and Nitrogenase Synthesis and Activity

Harry Christman,[1] Elsie Campbell,[1] Brett Phinney,[2] Wan-Ling Chiu,[3] and **John C. Meeks**[1*] (jcmeeks@ucdavis.edu)

[1]Dept. of Microbiology and [2]Genome Center, University of California, Davis, Calif.; and [3]Dept. of Biology, Virginia Commonwealth University, Richmond, Va.

Project Goals: Heterocyst-forming cyanobacteria, such as *Nostoc punctiforme*, are applicable for cost-effective biohydrogen production, providing the frequency of heterocysts in the filaments can be increased by genetic manipulations and the metabolic end product, H_2, is uncoupled from growth.

A rationale for this project is that heterocyst-forming cyanobacteria, such as *Nostoc punctiforme*, are applicable for cost-effective biohydrogen production, providing the frequency of heterocysts in the filaments can be increased by genetic manipulations and the metabolic end product, H_2, is uncoupled from growth. Heterocysts are the sites of nitrogenase and hydrogenase activities in these oxygenic photoautotrophs. Two keys to the genetic manipulation of heterocyst differentiation are detailed knowledge of the regulatory pathway for the initiation and maintenance of the approximately 10% heterocyst frequency in the free-living growth state, and identification of the regulatory targets of plant signals in establishment of a nitrogen-fixing symbiosis with *N. punctiforme*, wherein the heterocyst frequency increases to 30-40% of the cells and ammonium is released to the plant. The *N. punctiforme* symbiotic growth state is additionally characterized by an unbalanced metabolic physiology and a slow growth rate; these characteristics are also essential to duplicate for bioreactor H_2 production. Our working hypothesis is that heterocyst differentiation is modulated by a regulatory system operating in a cascade manner and that plant partners have evolved signals to co-opt the developmental regulatory pathway, as well as to control *Nostoc* growth and metabolism.

The experimental objectives are to apply transcriptomic and proteomic time course assays to wild-type and mutants of free-living cultures defective in stages of heterocyst differentiation in order to define epistatic relations. Expression patterns and potential protein modification of *N. punctiforme* in symbiotic growth with the bryophyte hornwort *Anthoceros punctatus* and the angiosperm *Gunnera manicata* will then be pursued.

Cluster analysis of 0.5 to 24 h time course experiments in free-living cultures have identified three distinct temporal clusters of 344 total up-regulated genes and one 28 member cluster of down-regulated genes. Comparisons of the cluster

members to steady state dinitrogen grown cultures indicate that by 24 h after induction of heterocyst differentiation, the cultures have not reached the steady state levels of gene expression. The results of this latter analysis allow us to identify four unique transcriptional patterns of stress related genes; 328 genes up-regulated following nitrogen starvation that decline to steady state values after 24 h, 147 genes that do not immediately change expression which are up-regulated after 24 h, 75 genes that do not change expression which are down-regulate after 24 h and 389 genes that are down-regulated following nitrogen starvation which are up-regulated after 24 h. The latter group of genes is enriched in those encoding core metabolic functions.

GTL

Thermotoga maritima Sugar Kinome

Irina Rodionva,[1] Xiaoqing Li,[1] Chen Yang,[1] Ying Zhang,[1] Ines Thiele,[2] Bernhard O. Palsson,[2] Dmitry Rodionov[1] (rodionov@burnham.org), and **Andrei Osterman**[1,3]* (osterman@burnham.org)

[1]Burnham Institute for Medical Research, La Jolla, Calif.; [2]Dept. of Bioengineering, University of California San Diego, La Jolla, Calif.; and [3]Fellowship for Interpretation of Genomes, Burr Ridge, Ill.

Project Goals: This project is focused on a systems-level understanding of biological hydrogen production using *Thermotoga maritima* as a model organism. We will address the basic science required to improve our understanding of hydrogen production from various carbon sources including glucose, cellulose, starch and xylan by this thermophilic microorganism. The overall goal is 1) to reconstruct the regulatory and metabolic network in *T. maritima* using various sets of "omics" data, 2) to integrate regulatory and metabolic networks into one "integrated" genome-scale model, 3) to confirm and validate the ability of the integrated model to predict processing of various environmental signals.

The marine hyperthermophilic bacterium *Thermotoga maritima* has extensive and highly diversified carbohydrate utilization machinery. A detailed reconstruction of this machinery including uptake mechanisms, biochemical transformations and transcriptional regulation is of key importance for the scope of our DOE-GTL-sponsored project "Systems-level understanding of hydrogen production by *Thermotoga maritima*" (see the accompanying poster by V. Portnoy et al.). Accurate functional assignment of carbohydrate utilization genes is challenging due to substantial variations of the respective pathways between species including frequent nonorthologous gene displacements and functionally divergent paralogs. Therefore, their homology-based annotations in various genomic databases are often incomplete and imprecise. To address this challenge we combine a subsystems-based approach to pathway analysis (implemented in the SEED genomic platform) with the experimental characterization of *signature* genes. In this study we applied this integrated approach to infer and experimentally assess substrate specificities within the *T. maritima* sugar kinome (SK) represented by at least 20 sugar kinases involved in a variety of carbohydrate utilization pathways. Sugar phosphorylation is an essential step in the overwhelming majority of such pathways. Whereas in *E. coli*, this step is often performed by uptake-associated phosphotransferases (PTS), in *T. maritima* it appears to be fully delegated to the members of its highly extended and diversified SK. Using genome context analysis (conserved operons and regulons) we were able to tentatively assign nearly all members of SK specific physiological roles in reconstructed pathways. We used a "matrix" approach to experimentally test these assignments and to explore the relationship between the inferred physiological roles and *in vitro* substrate preferences of respective enzymes. Purified recombinant proteins (obtained using expression strains from the Joint Center for Structural Genomics, www.jcsg.org) were tested for their kinase activity versus a panel of >40 different mono- and disaccharides. Remarkably, nearly all of the 15 experimentally characterized enzymes (from 4 structurally distinct superfamilies) displayed a strong preference towards a single physiological substrate. The results of this analysis provide the direct validation of reconstructed pathways in *T. maritima* and improve our ability to accurately annotate SK in other genomic and metagenomic datasets.

Most of the analyzed sugar utilization pathways are controlled by committed transcription factors (TF). *T. maritima* has an unusually large number of TFs of the so-called ROK-family that are composed of two domains, a N-terminal kinase-like effector domain and a C-terminal DNA-binding HTH-domain. We used a comparative genomic approach for the *ab initio* prediction of regulons tentatively controlled by all six members of this family. Five of them were implicated in the regulation of carbohydrate utilization pathways. Gel-shift mobility assays were used to confirm all of the predicted TF binding sites and to assess possible small molecule effectors, most of them, mono- or disaccharides involved in respective pathways. This study illustrates the power of the subsystems-based approach to comparative genomic reconstruction of metabolic and regulatory networks that will be further extended for the assessment of biohydrogen production by *T. maritima*.

GTL

Conservation and Variations in *Shewanella* Transcriptional Regulatory Network

Dmitry Rodionov[1]* (rodionov@burnham.org), Irina Rodionova,[1] Xiaoqing Li,[1] Pavel Novichkov,[2] Margaret Romine,[3] Mikhail Gelfand,[4] and **Andrei Osterman**[1,5] (osterman@burnham.org)

[1]Burnham Institute for Medical Research, La Jolla, Calif.; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [3]Pacific Northwest

National Laboratory, Richland, Wash.; [4]Institute for Information Transmission Problems RAS, Moscow, Russia; and [5]Fellowship for Interpretation of Genomes, Burr Ridge, Ill.

Project Goals: The project "Integrated Genome-Based studies of *Shewanella* Ecophysiology" is a component of the *Shewanella* Federation studies aimed at better understanding of the ecophysiology and speciation of this important genus.

Comparative genomics approach was used to infer transcriptional regulatory networks in *Shewanella oneidensis* MR-1 and fourteen other species of *Shewanella* with sequenced genomes. To accomplish this goal, we combined the identification of transcription factors (TF) and candidate TF-binding sites with the cross-genome comparison of regulons and the genome context analysis of candidate target genes. The identification of TF repertoire (termed regulome) in *Shewanella* points to the existence of between 113 and 199 regulons per genome. Overall, the pan-regulome of 15 *Shewanella* strains includes 66 groups of orthologous TFs present in all strains (the core regulome), 26 unique to single strains, and 242 present in two or more strains.

Using the combined comparative genomic approach we described 78 TF regulons that control the central metabolism, production of energy and biomass, metal ion homeostasis and stress response in the *Shewanella* lineage that split into two groups. TFs and respective DNA motifs in the first group of 37 regulons are conserved between *E. coli* and *Shewanella.* This group includes regulons associated with metabolism of nitrogen, amino acids, fatty acids, carbohydrates, and cofactors that are largely conserved among *Shewanella* spp. It also includes several global regulons (such as Crp), for which the conservation of respective regulatory sites is much weaker. Although some variations in the regulon content are observed between different members of the *Shewanella* group, the most striking differences in the overall regulatory strategy are revealed in comparison with *E. coli*. Among multiple trends in diversification and adaptive evolution of regulatory interactions in two lineages of gamma-proteobacteria are regulon "shrinking", "expansion", "mergers", and "split-ups". The second group of 41 regulons identified in *Shewanella* utilizes TFs that do not have orthologs in *E. coli*. Among them are novel regulons that are predicted to control degradation of branch chain amino acids (LiuR), fatty acids (PsrA), and sugar catabolism (NagR, ScrR, AraR, and BglR tentatively implicated in the control of utilization of N-acetylglucosamine, sucrose, arabinose and β-glucosides, respectively). Remarkably, a number of physiologically equivalent pathways in *E. coli* and *Shewanella* appear to be under control of nonorthologous TFs. For example, two regulators FruR and Crp that control a large number of catabolic genes in *E. coli* are functionally replaced in this capacity by distinct TFs, HexR and PdhR, in *Shewanella.* At the same time, the content and functional role of the Crp regulon in *Shewanella* is significantly shifted towards controlling the anaerobic respiration.

Experimental validation of novel predicted regulons in *Shewanella* species is underway. It includes combination of several approaches. We use gel-shift mobility assays to confirm TF-binding sites and effectors for selected novel TFs (e.g. NagR). The analysis of correlations between multiple microarray expression profiles provides additional information for the assessment of regulon predictions. We are planning to use systematic gene expression profiling in a panel of TF knockout mutants for validation and further refinement of the genomic reconstruction of the transcriptional regulatory network in *Shewanella* spp. This project is a component of the *Shewanella* Federation studies aimed at better understanding of the ecophysiology and speciation of this important genus.

GTL

Development of *Cyanothece* as a New Model Organism for Biological Hydrogen Production

Anindita Banerjee,[1] Michelle Liberton,[1] Hanayo Sato,[1] Jana Stöckel,[1] Lawrence Page,[1] Rick Huang,[2] Yinjie Tang,[2] Hogtao Min,[3] Louis A. Sherman,[3] Stephen J. Callister,[4] Richard D. Smith,[4] and **Himadri B. Pakrasi**[1,2]*

[1]Depts. of Biology and [2]Energy, Environmental and Chemical Engineering, Washington University, St. Louis, Mo.; [3]Dept. of Biological Sciences, Purdue University, West Lafayette, Ind.; and [4]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: The objective of this proposal is to develop the cyanobacterium *Cyanothece* as a model organism for photobiological hydrogen production. Members of the genus *Cyanothece* are unicellular oxygenic prokaryotes with the ability to fix atmospheric nitrogen. Our long-term goal is to develop a deep understanding of the metabolism of these microbes as it pertains to H_2 evolution. Specifically, we will use genome sequencing, microarrays, proteomics, mutagenesis, biochemical analysis and physiology, all of which are encased in a systems biology framework.

This project is focused on developing *Cyanothece*, a cyanobacterium, as a new model microbe to study biological hydrogen production. Cyanobacteria are oxygen evolving photosynthetic prokaryotes that live under wide ranges of environments and are key organisms in the harvesting of solar energy at a global level. Among them, unicellular species such as *Cyanothece* have the additional ability of nitrogen fixation, a process that is exquisitely sensitive to oxygen. To accommodate such incompatible processes in a single cell, *Cyanothece* produces oxygen and stores carbon in the form of glycogen during the day, and subsequently creates an O_2-limited intracellular environment to perform oxygen-sensitive processes such as N_2-fixation and H_2 production during the night. Thus, *Cyanothece* cells are natural bioreactors for the storage of captured solar energy with subsequent utilization at a different time during a diurnal cycle.

Our long-term goal is to gain a broad understanding of the capacity of hydrogen production by *Cyanothece* cells, using a systems biology approach. We are studying seven *Cyanothece*

strains with sequenced genomes and choose one with the best metabolic properties as the model organism. Our objective is to develop a two-stage photobiological hydrogen production process. The first stage is the production of biomass (glycogen) during photosynthesis (day time). In the second stage, H_2 will be produced from glycogen during O_2-limited dark fermentation (night time). We plan to use highly controlled and monitored cultivation methods combined with advanced analytical techniques to optimize both processes. Furthermore we will use genome sequence-based insights and advanced omics-level tools to optimize H_2-production during O_2-limited dark fermentation. These studies are expected to provide potential solutions to many of the existing challenges in hydrogen production by photosynthetic microbes, including the strategy for periodic production of an oxygen-limited environment in an oxygen-evolving cell under aerobic conditions

GTL

Systems-Level Understanding of Hydrogen Production by *Thermotoga maritima*

Vasiliy Portnoy[1]* (vasiliy.portnoy@gmail.com), Karsten Zengler,[1] Ines Thiele,[1] Ying Zhang,[2] Andrei Osterman,[2] and **Bernhard O. Palsson**[1] (palsson@ucsd.edu)

[1]Dept. of Bioengineering, University of California San Diego and [2]Burnham Institute for Medical Research, La Jolla, Calif.

Project Goals: The project addresses the basic science required to improve our understanding of hydrogen production from various carbon sources including glucose, cellulose, starch and xylan by *Thermotoga maritima*. The goals of the project are 1) to reconstruct the regulatory and metabolic network in *T. maritima* using various sets of "omics" data, 2) to integrate regulatory and metabolic networks into one "integrated" genome-scale model, 3) to confirm and validate the ability of the integrated model to predict processing of various environmental signals and hydrogen production. This project combines the state of the art computational and experimental procedures and will significantly advance our systems-level understanding of hydrogen production in *T. maritima* through the generation of new data sets, and their integration with existing knowledge in the form of a genome-scale network reconstruction that can be used to compute physiological functions and states in hydrogen producing organisms.

This project is focused on a systems-level understanding of biological hydrogen production using *Thermotoga maritima* as a model organism. We will address the basic science required to improve our understanding of hydrogen production from various carbon sources including glucose, cellulose, starch and xylan by this thermophilic microorganism. The overall goal is 1) to reconstruct the regulatory and metabolic network in *T. maritima* using various sets of "omics" data, 2) to integrate regulatory and metabolic networks into one "integrated" genome-scale model, 3) to confirm and validate the ability of the integrated model to predict processing of various environmental signals.

The current metabolic reconstruction of *T. maritima* contains 479 metabolic genes, 565 metabolites (non-unique) and 646 internal and external metabolic reactions. About 93% of these reactions (519) could be associated with a gene product. The total of 479 genes considered in this reconstruction corresponds to 25% of *T. maritima*'s genome. This reconstruction is based on the integration of the data from the literature supplemented by the extensive use of genomic techniques including a subsystem-based approach implemented in the SEED database. This high gene coverage also illustrates the amount of information being available for major metabolic pathways in *T. maritima*. The *T. maritima*'s metabolic reconstruction covers reactions and pathways of central metabolism, amino acid, nucleotide, lipid, and carbohydrate metabolism. The reconstruction accounts for the metabolism of 45 carbohydrates including some highly complex polysaccharides. Furthermore, the *T. maritima* model is able to produce hydrogen using various carbon sources.

The overall scope, content and quality is comparable with organisms for which more publications are available. Partially, this stage of development can be contributed to the availability of structural and detailed physiological studies. In fact, this reconstruction is the first that used extensively structural information for reaction/gene evidence during the reconstruction process. For example, crystallography data provided annotation evidence in some cases, confirming the function of the annotated gene as well as the candidate substrate and coenzyme utilization specific for the *T. maritima* enzyme.

Building on this metabolic reconstruction, we intend to identify genes and therefore reactions that are "missing" using competitive genomics approaches and computational algorithms. In order to improve the predictive abilities of the current model, we will also re-define the current biomass function by directly measuring the cellular composition of *T. maritima*. We will further reconstruct the transcriptional regulatory network (TRN) of *T. maritima*. TRN reconstruction will be based on a computational approach as well as on experimental methods using chromatin immunoprecipitation combined with whole-genome microarrays (ChIP-chip) or sequencing (ChIP-seq) to experimentally define genome-wide binding patterns of RNA polymerase, major sigma factors, and selected transcription factors in *T. maritima.* These experiments will be complemented with high-resolution gene expression profiles using tiled microarrays. The most important bioinformatics predictions of novel genes and pathways will be supported by focused validation experiments. We will then integrate the metabolic and regulatory networks to generate a comprehensive model of *T. maritima* that will allow for prediction of optimal production of hydrogen. Some examples illustrating our integrated approach to genomic reconstruction and experimental validation of metabolic pathways and regulatory circuits related to carbohydrate utilization are provided in the accompanying poster "Sugar kinome of *T. maritima*".

GTL

Systems-Level Kinetic Flux Profiling Elucidates a Bifurcated TCA Cycle in *C. acetobutylicum*

Daniel Amador-Noguez* (damador@princeton.edu), Jing Fan (jingfan@princeton.edu), Nathaniel Roquet (nroquet@Princeton.EDU), Xiao-Jiang Feng, Herschel Rabitz, and **Joshua D. Rabinowitz** (joshr@princeton.edu)

Dept. of Chemistry and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, N.J.

Project Goals: Microbial hydrogen (H_2) production holds great promise as a source of renewab e clean energy. A critical step towards more efficient biohydrogen production is improved understanding of the regulation of H_2-related metabolism. While a diverse range of microorganisms are capable of producing H_2, the complete intracellular pathways involved in H_2 production are known for only a few of them, and the full regulation of these pathways for none Advances in systems biology, analytical chemistry, and computer science are beginning to provide the tools required for effective quantitative modeling of complex biological networks. A critical goal is to develop models that are sufficiently accurate to enable rational control of the network behavior. The cellular metabolic pathways for H_2 production form a biological network whose rational control would have profound value. With the long term aim of enabling such control, we propose to develop integrated experimental-computational technologies for quantitative dissection of microbial hydrogen-producing metabolism. These tools will be broadly applicable to many microbial H_2 producers. We plan to illustrate them with two organisms: *Clostridium acetobutylicum* (possessing the fastest and highest yielding hexose fermentation pathway to H_2 of any microbe yet reported), and a novel class of thermophilic cyanobacteria that lack hydrogenase and produce H_2 at 62˚C via a nitrogenase-dependent pathway (*Synechococcus* species).

Clostridium acetobutylicum is an organism with great potential for the commercial production of butanol and hydrogen. Much work has been done to elucidate the metabolic pathways by which solvents are produced in this anaerobic bacterium. However, there are still key pathways of primary metabolism that remain unresolved, including the TCA cycle and amino acid biosynthesis.

Annotation of the genome sequence of *C. acetobutylicum* failed to identify obvious homologues of the enzymes of the right-side of the TCA cycle: citrate synthase, aconitase, and isocitrate dehydrogenase. These enzymes are traditionally required to synthesize α-ketoglutarate, the carbon skeleton of the glutamate-family of amino acids. To explain *C. acetobutylicum's* ability to synthesize glutamate and grow on minimal media, it was proposed that the TCA cycle could function in the reductive (counterclockwise) direction to produce α-ketoglutarate. Alternatively, it was suggested that glutamate is synthesized from ornithine by the arginine biosynthesis pathway running in reverse.

To elucidate the actual pathway that leads to α-ketoglutarate and glutamate, we studied the incorporation of various isotope-labeled nutrients into intracellular metabolites using liquid chromatography-tandem mass spectrometry (LC-MS/MS). These studies quickly ruled out both of the above hypotheses: feeding U-^{13}C-glucose labeled both succinate and α-ketoglutarate with similar kinetics, but the labeling patterns were inconsistent with carbon exchange between the compounds; feeding U-^{13}C ornithine labeled downstream arginine pathway compounds but not upstream ones or glutamate.

The actual route of α-ketoglutarate production was suggested by the observation that (despite the putative lack of citrate synthase), citrate was labeled faster than α-ketoglutarate upon feeding either U-13C-glucose or U-13C-acetate, with the labeling patterns of citrate, aconitate, and α-ketoglutarate consistent with turning of the right side of the TCA cycle in the oxidative direction. Interestingly, there was no passage of carbon from acetate into succinate, confirming the bifurcation of the TCA cycle.

In most organisms, the pro-chiral center of citrate has *S* stereochemistry. The position of ^{13}C atoms in glutamate and proline obtained from the U-^{13}C glucose and U-^{13}C acetate labeling experiments indicated, however, the production of citrate with *R* stereochemistry at the prochiral center. A homology search revealed a gene in *C. acetobutylicum* with ~60% sequence similarity to the *Re*-citrate synthase present in *C. kluyveri*. Efforts are ongoing to directly demonstrate the citrate synthase activity of the encoded protein biochemically.

Our results demonstrate that *C. acetobutylicum* has a bifurcated TCA cycle where α-ketoglutarate is produced in the oxidative direction from oxaloacetate and acetyl-CoA via citrate, and succinate is produced in the reductive direction from oxaloacetate via malate and fumarate. This observation is essential for construction of an accurate genome-scale model of *C. acetobutylicum* metabolism and lays the groundwork for better understanding of integration of biosynthetic metabolism with solvent and hydrogen production.

GTL

Systems Approach to Probing Hydrogen Regulation (SAPHyRe): The SurR Redox-Switched Transcriptional Regulator Controlling Hydrogen Production in *Pyrococcus furiosus*

Gina L. Lipscomb,[1]* Hua Yang,[1,3] Annette Keese,[2] Gerrit J. Schut,[1] B.C. Wang,[1] Michael Thomm,[2] Janet Westpheling,[1] Michael W.W. Adams,[1] and **Robert A. Scott**[1] (rscott@uga.edu)

[1]Dept. of Biochemistry & Molecular Biology, University of Georgia, Athens, Ga.; [2]Dept. of Microbiology, University of Regensburg, Germany; and [3]Centers for Disease Control and Prevention, Atlanta, Ga.

Project Goals: The goal of SAPHyRe (Systems Approach to Probing Hydrogen Regulation) is to develop a detailed systems-level description of the regulatory and metabolic networks controlling hydrogen production in the hyperthermophilic archaeon *Pyrococcus furiosus* (Pf). Pf will be used as the model organism to investigate its response to various environmental conditions relevant to all hydrogen-producing microorganisms, such as carbon and nitrogen sources, metal availability, and oxidative and reductive stresses. The outcome of this project will serve two purposes: 1) it will bring us one step closer to utilizing *P. furiosus* in development of H_2 as an alternative energy source and 2) it will serve as a model methodology for investigating the regulatory pathways of hydrogen production in other organisms.

Pyrococcus furiosus has the potential to produce hydrogen (H_2) efficiently from biomass, making it an excellent target organism for development of this alternative energy source. SAPHyRe (Systems Approach to Probing Hydrogen Regulation) explores the regulatory and metabolic networks controlling H_2 production in *P. furiosus* and will provide a model for investigating such networks in other H_2-producing organisms. A major control point for H_2 production in *P. furiosus* involves sulfur (S^0), which when present, causes a shift in metabolism from production of H_2 to H_2S. A key transcriptional regulator involved in this shift is the S^0 Response Regulator, SurR, which activates the hydrogenase operons along with related genes and represses genes involved in S^0 metabolism such as the recently described NAD(P)H Sulfur Reductase. SurR was discovered by affinity capture from cell extract with a bait DNA promoter region of the membrane-bound hydrogenase operon which is significantly down-regulated during the primary response to S^0. Biophysical, functional and structural characterization of SurR led to the discovery of an internal disulfide switch that controls its DNA-binding activity. The switch is an imbedded CxxC motif within the N-terminal helix-turn-helix DNA-binding domain, which when in the reduced thiol state, causes SurR to bind DNA in a sequence-specific manner. In the presence of S^0, the CxxC motif becomes oxidized to a disulfide, and SurR can no longer exert transcriptional control, resulting in deactivation of hydrogenase operons and derepression of S^0-metabolizing genes. This exemplary regulatory system mediating the interplay of H_2 and S^0 metabolism will be used in SAPHyRe as a guide to further our understanding of how H_2 production is affected by other key factors such as carbon and nitrogen sources, metal availability, and oxidative and reductive stresses.

GTL

Filling Knowledge Gaps in Biological Networks: Integrated Global Approaches to Understand H_2 Metabolism in *Chlamydomonas reinhardtii*

Christopher H. Chang,[1] Alexandra Dubini,[1] Florence Mus,[2] Venkataramanan Subramanian,[1,3] David M. Alber,[1] Peter A. Graf,[1] Glenn Murray,[3] Kwiseon Kim,[1] Arthur R. Grossman,[2] Matthew C. Posewitz,[3] and **Michael Seibert**[1]* (mike_seibert@nrel.gov)

[1]National Renewable Energy Laboratory, Golden, Colo.; [2]Carnegie Institution of Washington, Stanford, Calif.; and [3]Colorado School of Mines, Golden, Colo.

Project Goals: Development of photobiological H_2-production processes, a key component in pursuit of DOE's renewable energy mission, would be substantially accelerated by increasing our understanding of the extremely complex underlying biology. Although systems biology has evolved rapidly in recent years, the lack of comprehensive experimental data for a given organism prevents reliable predictive modeling based on biophysical representations. We will therefore employ petascale computing to address this issue by (1) computational parameter estimation to delimit the space of stable solutions for experimentally constrained metabolic models and (2) de novo numerical experimentation to characterize network capabilities under assumed but easily reconfigured kinetic models. The response will be characterized at the level of enzyme kinetic differential equation parameters. Through development of scalable software tools, iterative model building, and incorporation of experimental constraints generated by high-throughput "omics" technologies, a model of metabolism linked to H_2 production in the green alga, *Chlamydomonas reinhardtii*, will be constructed and tested biologically by systems biology approaches. Once a set of acceptable kinetic parameters has been computed, the model will then be used for high-performance optimization of H_2 output in the space of enzyme expression levels, subject to limitations on cell viability. Integration of a graphical job configuration interface into the popular Systems Biology Workbench will make our tools accessible to a broader user community. The work is envisioned as an important contribution toward long-term development of a complete in silico cell.

The goal of this project, jointly funded by the DOE GTL and SciDAC Programs, is to develop the means to globally map *in silico* all biological pathways in *Chlamydomonas reinhardtii* that can affect H_2 production by the organism. *C. reinhardtii* was the first alga with a sequenced genome (released in 2003 by the JGI), and it has recently emerged as a prototype organism for investigating fermentative processes and their regulation. *Chlamydomonas* has a complex anaerobic metabolic network that can be induced under dark, anaerobic conditions, where it can produce H_2 along with other fermentation products such as formate, acetate, ethanol, and CO_2. Previous studies focused on determining genes (by microarray and RT-PCR) that were differentially regulated as the result of shifting cultures of the parental strain CC-425 from aerobic growth to dark, anaerobiosis [1]. Indeed, anoxia led to differential expression of genes involved in fermentation and more specifically upregulation of the pyruvate formate lyase (PFL1) and pyruvate:ferredoxin oxidoreductase (PFR1) genes. Moreover, *C. reinhardtii* synthesized formate, acetate, and ethanol in the ratio 1:1:0.5. More recent experiments showed that in a strain lacking active hydrogenases, the ratio shifted to 2:1:1, respectively, and that succinate was produce instead of H_2. Interestingly the levels of transcripts encoding several proteins involved in fermentation also changed. Under this condition PFL1 was more upregulated and PFR1 was downregulated relative to the wild-type strain [2]. Together these results allowed us to establish a fermentative pathway model that provides information on the metabolic flexibility of the organism and H_2 metabolism.

In order to apply this and other experimental results discussed below, we have developed a computational model that integrates metabolic parameters obtained experimentally. We have encoded central carbon metabolism (glycolysis and the tricarboxylic acid cycle), oxidative phosphorylation, and fermentative metabolism known to occur in *C. reinhardtii* in the community-standard Systems Biology Markup Language. The model spans the cytosol; the mitochondrial intermembrane space, membranes, and matrix; and the chloroplast stroma. More explicitly, the modeling is based on a thermodynamic/kinetic paradigm, in which we seek to incorporate basic physical constraints and experimentally observed kinetic relationships directly. Merging of component pathway models, parsing of the merged model to C++ encoding the relevant ordinary differential equations, compiling, and linking to form a model-specific, high-performance executable program was accomplished through our software package, the *High-Performance Systems Biology Toolkit* (HiPer SBTK) [3]. A simple "make" process results in an executable program capable of simultaneous sampling and optimizing metabolite levels, fluxes, or sensitivities within the space of kinetic parameters or enzyme concentrations, with good scaling observed to 128 processing cores so far. Both local and global optimization is possible. Job configuration is possible through an auxiliary graphical interface or by direct editing of simple text. Stable model integration depends critically on the initial guess for metabolite levels, which in the absence of experimental data is generated by small manual trials. Key developments going forward include partial automation of initial guess generation, model expansion to include starch metabolism and photosynthetic light and dark reactions, standardized file format output for advanced visualization, and simultaneous sampling, fitting, and extremization of metabolites, fluxes, and sensitivities.

Our current biological research, which is supporting the computational studies, is focused on understanding the regulation of relevant pathways using proteomics-based approaches. Parental *C. reinhardtii* strain CC-425 was again grown under conditions where the cells were shifted from light-aerobic growth to dark-anoxic, H_2-producing conditions. Whole-cell protein extracts from these cultures were separated on 2D gels. The first dimension (isoelectric focusing) used a 24-cm, non-linear Immobiline pH Gradient (IPG) strip (pH 3-10), and the second dimension used a 12-14% polyacrylamide gradient gel. Proteins were then visualized using SYPRO Ruby staining, and analyzed with ImageMaster 2D Platinum 6.0 software to identify proteins differentially regulated under the two conditions. Preliminary analysis revealed that approximately 500 proteins were expressed under dark, anoxic conditions, compared to 600 under light, oxic condition. Further analysis to identify the differentially expressed proteins is in progress using a Micromass Quantitative-Time-of-Flight (Q-Tof) mass spectrometer.

In summary, high-throughput 'omics' techniques are examining the metabolic flexibility of *Chlamydomonas* and are being used to build metabolic computational models of the alga. Information obtained from the *in silico* modelling is expected to improve understanding of complexity of the metabolic network system, how cells adjust to changes in metabolite fluxes, what happens when specific metabolic reactions are blocked, and how the organism might be engineered to improve H_2-production yields.

References

1. Mus, F., Dubini, A., Seibert M., Posewitz M.C., and Grossman A.R. (2007) "Anaerobic acclimation in *Chlamydomonas reinhardtii*: Anoxic gene expression, hydrogenase induction and metabolic pathways", *J. Biol. Chem.* **282 (35)**, 25475-25486.
2. Dubini, A., Mus, F., Seibert M., Grossman, A.R., and Posewitz M.C., (2008) "Flexibility in anaerobic metabolism as revealed in a mutant of *Chlamydomonas reinhardtii* lacking hydrogenase activity ", *J. Biol. Chem. in press.*
3. Chang, C.H., Graf, P., Alber, D.M., Kim, K., Murray, G., Posewitz, M., and Seibert M., (2008) "Photons, Photosynthesis, and High-Performance Computing: Challenges, Progress, and Promise of Modeling Metabolism in Green Algae", *J. Phys. Conf. Ser.* **125**: 012048.

GTL

Examining the Molecular Basis for the Utilization of Alternative Redox Systems to Maximize Hydrogen Production in RubisCO-Compromised Mutants of Nonsulfur Purple Bacteria

Rick A. Laguna* (laguna.2@osu.edu) and **F. Robert Tabita**

Dept. of Microbiology, Ohio State University, Columbus, Ohio

Project Goals: The use of hydrogen as an energy source is attractive because the end-product is water, as opposed to carbon dioxide, a green house gas generated by the burning of fossil fuels. Hydrogen is normally produced by nonsulfur purple NSP) photosynthetic bacteria during nitrogen limiting conditions by the nitrogenase complex. However, the presence of ammonia in the environment normally repressed this process. We have shown that NSP photosynthetic bacteria possess an array of metabolic and regulatory capabilities that allow for the utilization of alternative redox sinks when the primary electron sink, carbon dioxide, is nullified via the inactivation or deletion of the RubisCO genes. These RubisCO-compromised mutant strains are able to derepress the synthesis of the nitrogenase complex under normal repressive conditions, thus allowing for the production of copious quantities of hydrogen gas. Such gain-of-function adaptive mutant strains have been obtained from *Rhodobacter capsulatus*, *Rhodobacter sphaeroides*, *Rhodospirillum rubrum*, and *Rhodopseudomonas palustris*. We are in the process of gaining a greater understanding of the molecular basis that allows for the utilization of the nitrogenase complex under normal repressive conditions, in order to maximize hydrogen production in RubisCO-compromised mutants of nonsulfur purple bacteria.

The Calvin-Benson-Bassham (CBB) reductive pentose phosphate pathway is responsible for the incorporation of CO_2 into cellular carbon under autotrophic growth conditions. Under photoheterotrophic conditions, CO_2 is primarily used as an electron acceptor via the CBB cycle in order to maintain redox poise within the cell. The key enzyme of the CBB pathway, RubisCO, catalyzes the actual CO_2 reduction step. Over the years (1-4), we have shown that nonsulfur purple (NSP) photosynthetic bacteria possess an array of metabolic and regulatory capabilities (5) that allow for the utilization of alternative redox sinks when the primary electron sink, CO_2, is nullified via the inactivation or deletion of the RubisCO genes. In order to grow photoheterotrophically, such RubisCO-compromised strains develop interesting strategies and alter their basic metabolic profile. For example, in many instances the derepression of nitrogenase synthesis occurs under normal repressive conditions. Such gain-of-function adaptive mutant strains have been obtained from *Rhodobacter capsulatus, Rhodobacter sphaeroides, Rhodospirillum rubrum*, and *Rhodopseudomonas palustris,* whereby such strains balance their redox potential via nitrogenase-catalyzed reduction of protons to hydrogen gas (1-4). Moreover, over the years we have shown that nitrogenase-derepressed mutant strains produce copious quantities of hydrogen gas by virtue of using the nitrogenase enzyme complex exclusively as a hydrogenase. We now show that knocking out competing redox balancing processes such as the CBB pathway plays a crucial role in maximizing hydrogen production in nitrogenase-derepressed strains.

We are in the process of gaining a greater understanding of the molecular basis that allows for the utilization of the nitrogenase complex under normal repressive conditions, particularly in RubisCO-compromised strains. A single point mutation in the nifA gene was shown to be important for nitrogenase derepression in RubisCO-compromised mutant strains of *R. capsulatus, R. sphaeroides*, and *R. palustris*. NifA is a key transcriptional activator of the structural genes encoding the nitrogenase complex (*nif*HDK). While current experiments suggest that a mutant NifA protein appeared to be responsible for derepression of the nitrogenase complex in R. *palustris*, in *R. sphaeroides* an additional thus far unidentified mutation appears to be involved in the derepression of the nitrogenase complex. Interestingly, no such *nif*A mutation was found in a nitrogenase-derepressed strain of a RubisCO knockout strain of *R. rubrum*. These results suggest that these organisms utilize different mechanisms to derepress nitrogenase synthesis in CBB-compromised mutant strains. Moreover, derepression of the nitrogenase complex is not the only way in which these organisms may balance their redox potential in the absence of a functional CBB pathway; e.g., some strains may derepress the synthesis of a novel sulfate reduction pathway or use other means to dissipate excess reducing equivalents. In all cases, the alternative redox balancing pathways are under exquisite control and are not maximally expressed unless the CBB pathway is knocked out, suggesting regulatory cross talk between these key pathways.

References

1. Joshi, H. M., and F. R. Tabita. A global two component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation, and nitrogen fixation. *Proc. Natl. Acad. Sci. USA.* 93 (1996), 14515–14520
2. Tichi, M. A., and Tabita, F.R. Maintenance and control of redox poise in *Rhodobacter capsulatus* strains deficient in the Calvin-Benson-Bassham pathway. *Arch. Microbiol.* 174 (2000) 322-333.
3. Tichi, M. A. and Tabita, F.R. Interactive control of *Rhodobacter capsulatus* redox balancing systems during phototrophic metabolism. *J. Bacteriol.* 183 (2001) 6344-6354.
4. Dubbs, J. M., and Tabita, F. R. Regulators of nonsulfur purple phototrophic bacteria and the interactive control of CO_2 assimilation, nitrogen fixation, hydrogen metabolism and energy generation. *FEMS Microbiol. Rev.* 28 (2004) 353-376.
5. Romagnoli, S., and Tabita, F. R. Carbon dioxide metabolism and its regulation in nonsulfur purple photosynthetic bacteria. In, Hunter C.N., Daldal F., Thurnauer M.C., and Beatty J. T. (2008) *The Purple Phototrophic Bacteria. Advances in Photosynthesis and Respiration*, Vol. 28, Springer, Dordrecht, The Netherlands, pp. 563-576

GTL

Pathways and Regulatory Network of Hydrogen Production from Cellulose by *Clostridium thermocellum*

Jonathan Millen[1]* (jonathan_millen@hotmail.com), Matthew Heckman,[1] Hsin-Yi Tsai,[1] Pin-Ching Maness,[2] Mary Lipton,[3] and **J.H. David Wu**[1] (davidwu@che.rochester.edu)

[1]University of Rochester, Rochester, N.Y.; [2]National Renewable Energy Laboratory, Golden, Colo.; and [3]Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: The overall objective of this research is to understand the H_2 metabolic pathway in this cellulose-degrading bacterium and its regulatory network at the molecular and systems levels. We hypothesize that: multiple hydrogenases work concertedly to contribute to hydrogen evolution; transcription factors control hydrogenase expression at the transcription level; and hydrogen metabolism is coupled to cellulolysis and fermentative pathways in this bacterium.

Clostridium thermocellum, a thermophilic, ethanogenic, and cellulolytic anaerobe, produces a complex extracellular cellulolytic organelle called the cellulosome. The cellulsome contains various depolymerizing enzymes that are arrayed on a protein scaffold and effectively degrades complex cellulosic substrates. During cellulose fermentation, the bacterium evolves hydrogen at a high rate. Analysis of its genome sequence reveals the existence of at least four putative hydrogenase genes central to hydrogen metabolism. Furthermore, at least 20 genes are potentially related to hydrogen metabolism. The bacterium is thus remarkably versatile in employing various enzymes, some of which are potentially novel, for hydrogen metabolism, indicating the significance of this biological process. Yet little is known concerning the pathway for hydrogen production and the regulatory mechanism/network that control these hydrogenase and the related genes as well as cellulolytic process and other metabolic pathways in the organism.

The overall objective of this research is to understand the H_2 metabolic pathway in this cellulose-degrading bacterium and its regulatory network at the molecular and systems levels. We hypothesize that: multiple hydrogenases work concertedly to contribute to hydrogen evolution; transcription factors control hydrogenase expression at the transcription level; and hydrogen metabolism is coupled to cellulolysis and fermentative pathways in this bacterium.

We are determining the hydrogenase expression and metabolic network nodes employing Real-Time RT-PCR, DNA microarray, and proteomic analyses on cells subjected to different culture conditions and metabolic pathway inhibitors, to probe differential expression of the various hydrogenases and their interrelationship with other cellular metabolic pathways. EMSA (electrophoretic mobility shift assay) coupled with mass spectrometry as well as the bacterial one-hybrid method are employed to identify transcription factors that bind to the promoters of the hydrogenase gene clusters. To extend quantitative analysis of genome expression to the protein level, we have initiated development of an Accurate Mass and Time (AMT) tag database for *C. thermocellum*. We have also established the methodology to profile system level changes in cellular protein expression in *C. thermocellum* using two-dimension separation and analysis on the ProteomeLab™ PF2D platform, a 2D chromatographic technique.

We will carry out ChIP-(chromatin immuno-precipitation-)on-chip assays targeting at transcription factors thus identified to map connections in the transcription factor network controlling linked metabolic pathways. Finally, we will purify FeFe-hydrogenases from its native producer and expressed recombinantly *E. coli* to determine their subunit compositions, endogenous redox partners, and the direction of reaction (hydrogen production vs. uptake) to shed light on their roles in hydrogen metabolism

The studies will provide important insights into the pathway and regulatory mechanism/network controlling hydrogen metabolism and cellulolysis as well as other pertinent metabolic pathways in this very intriguing cellulolytic and thermophilic bacterium, which catalyzes the rate-limiting cellulose-degradation reaction in a single-step process of biomass conversion (or CBP, Consolidated Bioprocessing). Detailed understanding of the pathway and regulatory mechanism/network will ultimately provide rationales for engineering, alternating, or deregulating the organism for biomass conversion.

GTL

Inference and Integration of Regulatory Dynamics in Metabolic Network Models

William J. Riehl[1]* (briehl@bu.edu), Hsuan-Chao Chiu[1]* (hcchiu@bu.edu), and **Daniel Segrè**[1,2] (dsegre@bu.edu)

[1]Graduate Program in Bioinformatics and [2]Depts. of Biology and Biomedical Engineering, Boston University, Boston, Mass.

http://prelude.bu.edu

Project Goals: The goal of this project is to develop new mathematical approaches for integrating regulatory information into metabolic network models. In addition to using experimental data for the inference of dynamically changing objective functions, we implement an ab initio approach to generate putative regulatory networks capable of optimal control.

Cellular metabolism dynamically adapts to changing environmental conditions and biological events. Maintenance of metabolic homeostasis is achieved through the regulation of metabolic enzymes in a complex interplay of transcriptional

and post-transcriptional mechanisms. In the context of genome-scale steady state models of metabolic networks, it has been shown that evolution may drive metabolic networks towards reaching computationally predictable optimal states, such as maximal growth capacity. However, the accuracy of these *in silico* models may vary significantly depending on environmental conditions, genetic perturbations, and the state of complex, and often unknown regulatory constraints. Many of these details, including the values of most kinetic parameters and the regulatory mechanisms that reprogram the metabolic pathways, have proven difficult to elucidate. Thus, understanding the principles that underlie metabolic regulation, and identifying approachable computational strategies for integrated modeling of metabolic and regulatory networks constitute an important open challenge, relevant for effectively modeling microbial systems towards increased production of valuable resources, such as biohydrogen.

Here, we propose two complementary approaches to this challenge, both extending current genome-scale steady-state models of metabolism. In the first approach [1], we develop a data-driven method based on linear optimization to infer time- or condition-dependent metabolic objectives, in the form of dynamically changing biomass compositions. Specifically, we seek to use experimental data (such as flux measurements, or mRNA expression levels) to infer best matching stoichiometrically balanced fluxes and metabolite sinks that maximally describe the metabolic objectives of multiple physiological states in an organism. We tested our approach by analyzing *E. coli* central carbon metabolism single gene mutants, and identified changes of metabolic objectives and potential compensation for reducing power in the pentose phosphate pathway. In the second, theory-driven approach [2], we develop a new mathematical method for predicting the regulatory network that controls a metabolic pathway, based on optimality criteria similar to the ones used in Flux Balance Analysis. The model hypothesizes that, once a steady-state metabolic system is perturbed, regulatory feedback based on the flux imbalance induced by the perturbation will gradually restore homeostasis. This feedback can either bring the system back to the original steady-state or, in a switch-like behavior, take it to a different state. This model has been applied to predicting optimal regulatory responses both in a simple toy network and in a fragment of the glycolysis pathway.

References

1. Hsuan-Chao Chiu and Daniel Segrè: Comparative determination of biomass composition in differentially active metabolic states, *Genome Informatics*, 2008, *In press*.
2. William J. Riehl and Daniel Segrè: Optimal metabolic regulation using a constraint-based model, *Genome Informatics*, 2008, *In press*.

GTL

Effects of Mutational Modification of Electron Pathways of *Desulfovibrio* Strains

Kimberly L. Keller[1,2]* (kellerkl@missouri.edu), Steve D. Brown,[3] Iris Porat,[3] and **Judy D. Wall**[1,2] (WallJ@missouri.edu)

[1]University of Missouri, Columbia, Mo.; [2]VIMSS (Virtual Institute for Microbial Stress and Survival), http://vimss.lbl.gov, Berkeley, Calif.; and [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: We propose to examine the capacity of the abundant soil anaerobes, sulfate-reducing bacteria, for hydrogen generation from organic acids. A comparison of *D. vulgaris* Hildenborough and *Desulfovibrio* G20 should provide insights into the limitations and potential of hydrogen production by members of this large group of strictly anaerobic bacteria. To explore this metabolic capacity, we will: 1) create a deletion of the gene encoding acetate kinase to confirm that this enzyme is essential for substrate-level phosphorylation during pyruvate fermentation, and determine the effect of this deletion on the efficiency of pyruvate and lactate respiration; 2) determine the enzyme(s) responsible for oxidizing pyruvate during fermentation and the role of formate, if any, in pyruvate fermentation; and 3) channel electrons from alternative sinks to hydrogen during fermentation and determine the effects of removal of those sinks on the fermentation efficiency. The total metabolism, flux through the pathways, and regulation are likely to be limiting factors which we can elucidate in our experiments.

Fermentative hydrogen generation provides a mechanism for anaerobic microbes to release electrons in a neutral fashion during oxidation of vast quantities of organic matter. We are examining the capacity of the abundant soil anaerobes, sulfate-reducing bacteria, for hydrogen generation from organic acids. These apparently simple pathways have yet to be clearly established. Because the tools for genetic and molecular manipulation of sulfate-reducing bacteria of the genus *Desulfovibrio* are functional, our efforts are focusing on two strains, *D. vulgaris* Hildenborough and *Desulfovibrio* G20. A markerless deletion protocol has already been established in *D. vulgaris* Hildenborough and preliminary experiments with G20 show that the counterselectable marker on which the protocol is based is quite effective. Since G20 is able to grow reproducibly in defined medium in the absence of sulfate as a terminal electron acceptor on a number of substrates, we are using microarray analysis to compare the changes in gene expression between wild-type G20 and various G20 mutants that may be affected in fermentation. We have already begun analysis of hydrogen production and metabolite changes of key G20 Tn*5* mutants (generously made available through the ESPP2 collaboration from A. Deutschbauer and A. Arkin). Ultimately, markerless deletions of any genes found to have an impact on hydrogen production will be constructed, allowing us to make sequen-

tial gene deletions to fully explore the metabolic processes that may channel reductant to hydrogen.

GTL

Modeling Electron Flow in *Rhodobacter sphaeroides* for the Identification of Potential Approaches to Maximize Hydrogen Production

L. Safak Yilmaz[1,*] (yilmaz@cae.wisc.edu), Alison P. Sanders, Rodolfo Perez,[1] Wayne S. Kontur,[2] Timothy J. Donohue[2] (tdonohue@bact.wisc.edu), and **Daniel R. Noguera**[1] (noguera@engr.wisc.edu)

[1]Dept. of Civil and Environmental Engineering and [2]Dept. of Bacteriology, University of Wisconsin, Madison, Wis.

Project Goals: To develop electron balance and metabolic network models for the prediction of potential electron sinks in *Rhodobacter sphaeroides* which can be modified to maximize solar-powered hydrogen production by this organism.

Introduction and Objective: The metabolically versatile organism *Rhodobacter sphaeroides* produces hydrogen while using light as an energy source and organic substrates as electron donor. This project aims at optimizing hydrogen production from *R. sphaeroides* by identifying the main electron sinks and genetically modifying the organisms to divert as high a fraction of electrons as possible to hydrogen generation.

Methodology: Cultures were grown in illuminated anaerobic batch reactors until stationary phase was reached. The overall electron flux to biomass, soluble microbial products (SMPs), and polyhydroxybutyrate (PHB) were quantified by chemical oxygen demand measurements of liquid and solid phases. Gas production was measured with respirometry and the hydrogen content of the biogas was determined by gas chromatography from endpoint headspace samples. Furthermore, a fuel cell connected to the respirometer was used as an additional hydrogen sensor.

Two models are being developed. First, an overall electron flow model is used to understand the global distribution of electrons, and second, a metabolic model of *R. sphaeroides* is being developed for detailed analyses of metabolic pathways and their potential contribution to hydrogen production. So far, a central metabolic network (CMN) that involves carbon metabolism, electron transport chain, and flow of electrons to hydrogen has been established. Computational flux balance analyses (FBA) have been performed to qualitatively test the model.

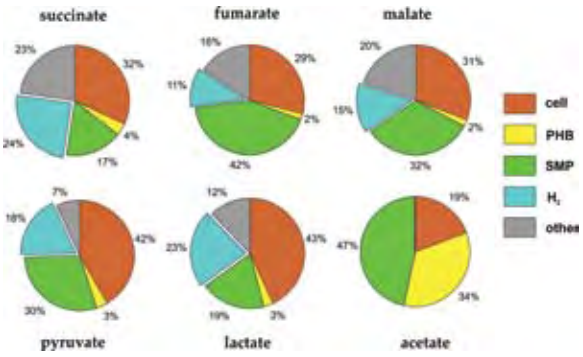


Figure 1. Distribution of substrate electrons in photosynthetically grown *R. sphaeroides*. Glutamate was used as a nitrogen source to promote hydrogen production in all tests. The carbon source was varied as indicated.

Results: The electron distribution from the growth of wild type cells with six different organic acids as carbon sources is given in Figure 1. Generally, an electron balance accounting for more than 80% of the reducing power was achieved. Hydrogen production efficiency was best from lactate and succinate with about a quarter of electrons flowing to hydrogen. Only acetate did not support detectable hydrogen production, possibly because of increased flow of electrons to PHB. The fraction of PHB was small for other substrates, while SMPs turned out to be an important electron sink in most experiments. Although hydrogen quantification was based on headspace measurements, the potential of using small-scale fuel cells as on-line sensors was tested. A linear relationship between gas production rate and voltage output in the fuel cell was observed in a wide range of gas rates (see Figure 2).

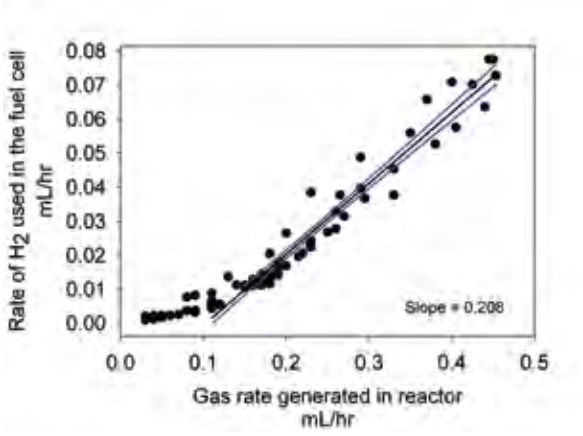


Figure 2. Conversion of biohydrogen to electricity in a proton exchange membrane-based fuel cell. The y-axis shows hydrogen used by the fuel cell, back-calculated from the voltage readout. The linear relationship suggests that hydrogen can be quantified by the fuel cell in a certain range of production rate. The slope indicates that around 20% of the hydrogen gas generated in the reactor was converted to electricity in the fuel cell.

The metabolic modeling effort has so far centered on creating the CMN model. At present, it correctly simulates growth (flux to biomass) coupled to hydrogen production

when an organic acid is used as the substrate, light is the energy source, and no external electron acceptors are provided. Consistent with experimental observations, the model predicts no hydrogen production when cells are grown under aerobic conditions in the absence of light.

Future Work: The electron balance modeling approach will continue to be used as a tool to quantify electron fate on wild type cultures under different growth conditions (e.g., different substrates and nitrogen sources and different growth stages) and on mutants generated to investigate the metabolic networks that impact hydrogen production. Experimental results will also be used to continue the development of the metabolic network model, which will be expanded to include metabolic networks necessary to simulate experimental observations with wild type and mutant strains.

SBIR Bioenergy Research

SBIR

Consolidated Bioelectrochemical Processing of Cellulosic Biomass to Ethanol and Hydrogen

T. Shimotori[1]* (shimo@mfctech.net), M. Nelson,[1] and H. May[2]

[1]MFC Technologies, LLC, Mount Pleasant, S.C.; and [2]Dept. of Microbiology and Immunology, Medical University of South Carolina, Charleston, S.C.

Project Goal: Enhance cellulosic ethanol fermentation through the use of an ethanol and electricity generating microbial consortium in a consolidated bioprocess.

Consolidated bioprocessing of cellulosic biomass, as originally proposed by Dr. Lynd, leverages the catalytic activity of cellulolytic and ethanologenic bacteria to produce ethanol.[1,2] Thermophilic bacteria such as *Clostridium thermocellum* and *Thermoanaerobacter thermosaccharolyticum* are typically used in this process due to their rapid and effective ability to metabolize cellulose and hemicellulose. However, the formation of by-products, especially fatty acids, is common. The by-product formation not only reduces the ethanol yield but also inhibits the overall fermentation. As a solution to this problem, Dr. Lynd's group recently reported a method to reduce by-product yields by genetic engineering of a cellulolytic bacterium.[3]

As an alternative way to tackle this problem, MFC Tech proposed to use a consolidated bioelectrochemical process to convert cellulosic biomass into ethanol or hydrogen. Electricigenic bacteria capable of consuming fatty acids are incubated with cellulolytic bacteria in thermophilic microbial fuel cells (MFCs), where fatty acids are converted to carbon dioxide and electricity is generated. MFCs have been extensively studied for years as summarized in Dr. Logan's review.[4] Only recently, however, a few research groups reported such thermophilic electricigenic bacteria.[5-7] Dr. May and colleagues at MUSC successfully enriched a mixed community of thermophilic electricigenic bacteria from marine sediment that consume acetate in MFCs at 60°C.[6] Analysis of the 16S rRNA genes revealed a community that includes bacteria most closely related to *Deferribacter* spp.[8-10] and Gram positive Firmicutes, particularly of *Thermincola* spp.[11,12] We used this mixed culture and an isolate of *Thermincola ferriacetica* as thermophilic electricigenic bacteria in combination with cellulolytic bacteria in cellulose-fed MFCs.

We operated air-cathode MFCs with the thermophilic electricigenic bacteria at 60°C. Once the MFCs started producing steady current using acetate, the anode chamber was filled with cellulose-containing media and inoculated with *C. thermocellum*. Although these MFCs produced ethanol and electricity, ethanol production was lower and acetate production was higher compared with cellulosic fermentation without MFC. This is because the reducing potential necessary for ethanol production was channeled to electricity production.

Due to the failure of a consolidated bioelectrochemical process in a MFC, we adapted it to a microbial electrolysis cell (MEC). A MEC, although similar to a MFC, is entirely anaerobic and can produce hydrogen from fatty acids with the help of added electricity. For a review of MECs, see Dr. Logan's recent report.[13] Consolidated bioelectrochemical processing of cellulose in thermophilic MECs using the electricigenic bacteria and *C. thermocellum* resulted in higher ethanol/acetate ratios compared with those by fermentation alone. An even better result was obtained when the electricigenic bacteria and *C. thermocellum* were co-cultured with *T. thermosaccharolyticum* in a MEC. We also demonstrated hydrogen production from cellulosic biomass in a similar system.

References

1. Lynd, L.R. *et al.* Consolidated bioprocessing of cellulosic biomass: an update. *Curr. Opin. Biotechnol.* **16**, 577-583 (2005).
2. Lynd, L.R. *et al.* Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506-577 (2002).
3. Shaw, A. J. *et al.* Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. *PNAS.* **105**, 13760-13774 (2008).
4. Logan, B.E. Microbial fuel cells. *Wiley-Interscience* (2008).
5. Jong, B. *et al.* Enrichment, performance and microbial diversity of a thermophilic mediatorless microbial fuel cell. Environ. Sci. Technol. **40**:6449-6454 (2006).
6. Mathis, B. J. *et al.* Electricity generation by thermophilic microorganisms from marine sediment. Appl. Microbiol. Biotechnol. **78**:147-155 (2008).
7. Wrighton, K. C. *et al.* A novel ecological role of the Firmicutes identified in thermophilic microbial fuel cells. ISME J. **2:**1146-1156 (2008).
8. Miroshnichenko, M. L. et al. *Deferribacter abyssi* sp. nov., an anaerobic thermophile from deep-sea hydrothermal

* Presenting author

vents of the Mid-Atlantic Ridge. *Inter. J. Syst. Evol. Microbiol.* **53**, 1637-1641 (2003).

9. Takai, K. *et al. Deferribacter desulfuricans* sp. nov., a novel sulfur-, nitrate- and arsenate-reducing thermophile isolated from a deep-sea hydrothermal vent. *Inter. J. Syst. Evol. Microbiol.* **53**, 839-846 (2003).
10. Greene, A. C. *et al. Deferribacter thermophilus* gen. nov., sp. nov., a novel thermophilic manganese- and iron-reducing bacterium isolated from a petroleum reservoir. *Int. J. Syst. Bact.* **47**, 505-509 (1997).
11. Sokolova, T. G. *et al. Thermincola carboxydiphila* gen. nov., sp. nov., a novel anaerobic, carboxydotrophic, hydrogenogenic bacterium from a hot spring of the Lake Baikal area. *Int. J. Syst. Evol. Microbiol.* **55**, 2069-2073 (2005).
12. Zavarzina, D. G. *et al. Thermincola ferriacetica* sp. nov., a new anaerobic, thermophilic, facultatively chemolithoautotrophic bacterium capable of dissimilatory Fe(III) reduction. *Extremophiles* **11**, 1-7 (2007).
13. Logan, B. E. *et al.* Microbial electrolysis cells for high yield hydrogen gas production from organic matter. *Environ. Sci. Technol.* **42**:8630-8640 (2008).

SBIR

Clostridium phytofermentans: Genome Sequence of a Model System for the Direct Conversion of Plant Biomass to Fuels

E. Petit,[1] T. Warnick,[1] W. Latouf,[1,3] A. Tolonen,[4] N. Ivanova,[5] A. Lykidis,[5] M. Land,[6] L. Hauser,[6] N. Kyrpides,[5] B. Henrissat,[7] G. Church,[4] D. Schnell,[2] S. Leschine,[1] and **J. Blanchard**[1]* (blanchard@microbio.umass.edu)

[1]Dept. of Microbiology and [2]Dept. of Biochemistry and Molecular Biology, University of Massachusetts, Amherst, Mass.; [3]Qteros Inc., Amherst, Mass.; [4]Dept. of Genetics, Harvard Medical School, Boston, Mass.; [5]DOE Joint Genome Institute, Genome Biology Program, Production Genomics Facility, Walnut Creek, Calif.; [6]Oak Ridge National Laboratory, Life Sciences Division, Oak Ridge, Tenn.; and [7]Architecture et Fonction des Macromolécules Biologiques, CNRS, and Universités d'Aix-Marseille I and II, France

Project Goals: Our long-term goal is to replace gasoline with fuels derived from ecologically and economically sustainable plant feedstocks. The objective of this project is to use an integrative systems level approach to characterize and manipulate the consortium of genes required to turn plant biomass into ethanol using *Clostridium phytofermentans*.

Clostridium phytofermentans produces biofuels directly from a broad range of industrially relevant feedstocks. Thereby, alleviating the need for a separate industrial process that first breaks down the cellulose components into sugars before fermenting them to biofuels. The *C. phytofermentans* genome contains; (1) more and a greater diversity of glycoside hydrolases than any known biofuel-related bacteria; (2) a cornucopia of transport systems for assimilating solublized mono- and oligomeric carbohydrates; (3) no evidence of dockerin, cohesin and/or scaffoldin-like proteins or other cellulosome features; (4) three proteinaceous microcompartment loci involved in alcohol production, and; (5) seven putative hydrogenase that may code for novel signaling and metabolic pathways. The influx of biofuels-related genes via horizontal transfer from other microbes may explain why *C. phytofermentans* is capable of fermenting all major carbohydrate components of biomass. These and other attributes make *C. phytofermentans* an excellent model organism for understanding the direct conversion of plant biomass to biofuels.

SBIR

A Genome-Wide Perspective on the Regulation of Plant Carbohydrate Conversion to Biofuels in *Clostridium phytofermentans*

E. Petit,* T. Warnick, S. Leschine, and **J. Blanchard** (blanchard@microbio.umass.edu)

Dept. of Microbiology, University of Massachusetts, Amherst, Mass.

Project Goals: Our long-term goal is to replace gasoline with fuels derived from ecologically and economically sustainable plant feedstocks. The objective of this project is to use an integrative systems level approach to characterize and manipulate the consortium of genes required to turn plant biomass into ethanol using *Clostridium phytofermentans*.

The economic costs of degrading cellulosic biomass currently hold back widespread use of plant biomass from agricultural and forestry wastes as a domestic renewable alternative to gasoline. *Clostridium phytofermentans* ferments all major component of the plant cell wall to ethanol and is an emerging model organism for understanding the direct conversion of plant biomass to fuels. The *C. phytofermentans* genome contains more carbohydrate degradation enzymes, more carbohydrate transporters and more transcriptional regulatory factors than any other biofuels-related microorganism. In order to build a foundation for engineering improvements on industrially relevant feedstocks, microarray experiments were carried out on a number of purified plant cell wall carbohydrates. The results demonstrate that *C. phytofermentans* regulates the stoichiometry of the plant degradative and assimilatory machinery in response to substrate availability. Because of the modularity of the sugar degradation and assimilation systems and the lack of a cellulosome, further improvement in biomass degradation through genetic manipulation is relatively straightforward.

Systems Environmental Microbiology

The Virtual Institute of Microbial Stress and Survival VIMSS:ESPP

GTL

ESPP Functional Genomics and Imaging Core (FGIC): Cell Wide Analysis of Metal-Reducing Bacteria

Aindrila Mukhopadhyay,[1,6]* Edward Baidoo,[1,6] Kelly Bender[5,6] (bender@micro.siu.edu), Peter Benke,[1,6] Swapnil Chhabra,[1,6] Elliot Drury,[3,6] Masood Hadi,[2,6] Zhili He,[4,6] Jay Keasling[1,6] (keasling@.berkeley.edu), Kimberly Keller,[3,6] Eric Luning,[1,6] Francesco Pingitore,[1,6] Alyssa Redding,[1,6] Jarrod Robertson,[3,6] Rajat Sapra,[2,6] Anup Singh[2,6] (aksingh@sandia.gov), Judy Wall[3,6] (wallj@missouri.edu), Grant Zane,[3,6] Aifen Zhou,[4,6] and Jizhong Zhou[4,6] (jzhou@rccc.ou.edu)

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Sandia National Laboratories, Livermore, Calif.; [3]University of Missouri, Columbia, Mo.; [4]University of Oklahoma, Norman, Okla.; [5]Southern Illinois University, Carbondale, Ill.; and [6]Virtual Institute of Stress and Survival (vimss.lbl.gov).

Project Goals: The primary goal of the Environmental Stress Pathway Project (ESPP) is a rigorous understanding of the sulfate reducing bacterium (SRB), *Desulfovibrio vulgaris* Hildenborough physiology and its ability to survive in its environment.

The primary goal of the Environmental Stress Pathway Project (ESPP) is a rigorous understanding of the sulfate reducing bacterium (SRB), *Desulfovibrio vulgaris* Hildenborough physiology and its ability to survive in its environment. This knowledge provides the basis for discerning the biogeochemistry at metal contaminated sites, for bioremediation and natural attenuation for toxic metals. The FGIC focuses on mapping these responses at a cell wide level using systems biology approaches. In the last one year, our methods that have been optimized and utilized over the years to study a variety of growth/ stress conditions and mutants, were extended to study more environmentally relevant physiological conditions and the interaction of *D. vulgaris* with other microbes.

Generating high quality biomass continues to be a critical aspect for all our functional genomics studies. This was found to be especially important for complex conditions such as biofilm formation, long term exposure to stress and the study of mixed cultures containing multiple organisms. These studies were conducted in close collaboration with the Applied and Environmental Core and additional methods development had to be undertaken to address many unique aspects of studying such complex systems. For example, the study of the *D. vulgaris* / *Methanococcus maripaludis* syntrophic co-culture (serving as a model of naturally occurring SRB/ Methanogen interactions) required optimization of microarray hybridization methods and an alternate workflow for iTRAQ proteomics application. Our team also has advanced tools for metabolite level analysis, such as a ^{13}C isotopomer based flux analysis which provides valuable information about bacterial physiology. However the study of individual organisms in a mixed culture using existing flux analysis methods is difficult since the method typically relies on amino acids from hydrolyzed proteins from a homogenous biomass. To overcome the need to separate the target organism in a mixed culture, we successfully explored the idea that a single highly-expressed protein could be used to analyze the isotopomer distribution of amino acids from one organism. An overview of there studies and key observations are presented.

Additionally we continued to collect cell wide data in *Shewanella oneidensis* and *Geobacter metallireducens* for comparative studies. Improved methods for extraction and high throughput of metabolite analysis using CE-MS and LC-MS were applied to several studies underway in ESPP, many of which are also required quantitation. In collaboration with Computational Core, a novel FTICR-MS based method for a comparative $^{12}C/^{13}C$ based metabolite analysis was also developed to enable a direct comparison of control and experimental samples for relative quantitation.

Improved methods for generating stable knockout mutants and marker-less mutants in *D. vulgaris* has been now been widely used to follow up hypothesis from a majority of our stress response and most valuably in confirming candidates for "missing steps" in *D. vulgaris* metabolism. Sets of targeted mutants are also being constructed to study the large number of two component signaling systems in *D. vulgaris*. To ensure a complete understanding of regulatory mechanisms, the study alternative regulatory mechanisms such as small non-coding RNAs are also underway. In collaboration with the computational core, work is in progress to set up searchable databases of all our large data sets, including proteomics, metabolite and flux data.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

* Presenting author

GTL

Analysis of a *Desulfovibrio vulgaris* Small RNA and Its Target Under Various Stress Conditions

Andrew S. Burns[1,2] and Kelly S. Bender[1,2]* (bender@micro.siu.edu)

[1]Microbiology Dept., Southern Illinois University, Carbondale, Ill.; and [2]VIMSS (Virtual Institute for Microbial Stress and Survival), http://vimss.lbl.gov

Project Goals: Our goal is to understand the intricate regulatory cascades involved in how the model metal reducer *Desulfovibrio vulgaris* and its relatives respond to stressors in contaminated sites.

One of the aims of the Environmental Stress Pathway Project is to elucidate regulatory networks critical to processes of interest to the DOE. As such, our goal is to understand the intricate regulatory cascades involved in how the model metal reducer *Desulfovibrio vulgaris* and its relatives respond to stressors in contaminated sites. One approach we are taking to meet this challenge is the identification and analysis of small non-coding RNA molecules (sRNAs). Ranging in size from 20-200 nucleotides, sRNAs predominantly affect gene regulation by binding to complementary mRNA in an anti-sense fashion and therefore provide an immediate regulatory response independent of protein modification. Here we report the analysis of Dv-sRNA2, a molecule previously identified from a random small RNA clone library, and its target gene DVU0678- a hypothetical protein only present in the *D. vulgaris* strains Hildenborough and DP4.

While expression of Dv-sRNA2 has been confirmed, its regulatory role is currently unknown. Dv-sRNA2 is located in the same chromosomal region as its putative target DVU0678, but on the opposite strand. As such, a Dv-sRNA2 deletion mutant cannot be constructed without affecting the expression of DVU0678. To circumvent this problem, the gene encoding Dv-sRNA2 was cloned into the stable vector pMO719 to elucidate the effects of over-expression. This construct resulted in a strain (KB100) containing two copies of Dv-sRNA2 under the control of their native promoter. Phenotypic analysis of the KB100 strain compared to a control strain harboring an empty vector indicated no difference in growth under normal 37°C/pH 7 and pH 6 growth conditions. However, a slight increase in growth rate was observed for KB100 when grown at 45°C as well as when 45°C growth was shifted to 50°C during early log-phase. This minimal phenotypic difference likely resulted from similar expression rates of Dv-sRNA2 in strain KB100 and the control. Since it appeared that the merodiploid strain was not over-expressing the sRNA gene, a strain was constructed in which the entire Dv-sRNA2/DVU0678 region was deleted via maker exchange. Analysis of this deletion strain (KB102) compared to the wild type indicated similar growth patterns during 25°C and 37°C growth at pH 7. However, increasing the growth temperature to 45°C increased the lag phase of KB102 compared to the wild type by nine hours. While these data implicate that Dv-sRNA2 and DVU0678 are involved in the *D. vulgaris* heat shock response, studies involving other stressors are currently underway. New strategies for constructing over-expression strains are also being developed to help predict the role individual sRNAs have on the physiology and transcriptional response of *D. vulgaris* under multiple environmental conditions.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Functional Characterization of Microbial Genomes by Tagged Transposon Mutagenesis

Adam Deutschbauer[1,4]* (AMDeutschbauer@lbl.gov), Julia Oh,[2] Morgan Price,[1,4] Jennifer Kuehl,[1,4] Paramvir Dehal,[1,4] Corey Nislow,[3] Ronald W. Davis,[2] and **Adam P. Arkin**[1,4]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Stanford Genome Technology Center, Stanford, Calif.; [3]University of Toronto, Toronto, Ontario, Canada; and [4]VIMSS (Virtual Institute for Microbial Stress and Survival) http://vimss.lbl.gov

Project Goals: A primary goal of the Environmental Stress Pathway Project (ESPP) is a systems-level model of sulfate-reducing bacteria (SRB) metabolism, stress responses, and gene regulation. However, current systems-level analyses of less studied bacteria such as SRBs are limited by the presence of numerous uncharacterized genes and an over reliance on annotations from well studied bacteria such as *E. coli*.

A primary goal of the Environmental Stress Pathway Project (ESPP) is a systems-level model of sulfate-reducing bacteria (SRB) metabolism, stress responses, and gene regulation. However, current systems-level analyses of less studied bacteria such as SRBs are limited by the presence of numerous uncharacterized genes and an over reliance on annotations from well studied bacteria such as *E. coli*. Therefore, it is imperative that rapid and quantitative methods are developed to determine microbial gene function in a high-throughput manner. To meet this challenge, we are developing a mutagenesis and phenotyping strategy that is comprehensive across the genome and applicable to any microorganism amenable to transposon mutagenesis. We have cloned and sequence-verified 4280 tag modules into a Gateway entry vector. Each tag module is a 175 base pair element containing two unique 20 base pair sequences, the UPTAG and DOWNTAG, flanked by common PCR priming sites. Each tag module can then be rapidly transferred *in vitro* to any DNA element, such as a transposon, that is made Gateway compatible. Transposon mutants

marked by the modules will be sequenced to determine which of the 4280 tag modules was used and which gene was disrupted. Transposon mutants can be rapidly re-arrayed into a single pool containing 4280 uniquely tagged, sequence-verified mutant strains. By sequencing saturating numbers of transposon mutants, we can identify and assay mutants in most nonessential genes in a given genome. The fitness of each mutant in the pool will be monitored by the hybridization of the barcodes to an Affymetrix microarray containing the tag complements in a system identical to that used for the yeast deletion collection. To facilitate both strain construction and mutant pool phenotyping, we have implemented a robotic infrastructure (both aerobic and anaerobic) for assay setup and automated mutant pool growth, collection, and processing. Here we describe the initial application of our approach to the metal-reducing bacterium *Shewanella oneidensis* MR1 and the SRBs *Desulfovibrio desulfuricans* G20 and *Desulfovibrio vulgaris* Hildenborough.

The successful completion of this project will enable the quantitative phenotypic analysis of thousands of mutants across a wide range of conditions. These data will be used to assign gene function on a global scale, aid in the identification of missing metabolic enzymes, provide insight into the functional connectivity of different pathways, and enable the construction of genome-wide models of SRB function and activity.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

The Development and Application of an Integrated Functional Genomics Platform in *Desulfovibrio desulfuricans* G20

Adam Deutschbauer,[1,2] Jennifer Kuehl[1,2]* (JVKuehl@lbl.gov), Morgan Price,[1,2] Paramvir Dehal,[1,2] Terry C. Hazen,[1,2] and **Adam P. Arkin**[1,2]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [2]VIMSS (Virtual Institute for Microbial Stress and Survival), http://vimss.lbl.gov

Project Goals: We are developing an integrated functional genomics platform in the SRB *Desulfovibrio desulfuricans* G20 (G20). The platform is compromised of phenotypic arrays, gene expression arrays, and a transposon mutant library. A hallmark of our transposon library is the ability to pool thousands of mutants for parallel phenotypic analysis. Profiling our transposon library under a large number of conditions will enable us to link a large fraction of the G20 genome to specific cellular phenotypes. Here we present our experimental approaches and preliminary data from each type of experiment.

The Environmental Stress Pathway Project (ESPP) focuses on the systems-biology and environmental activity of the sulfate-reducing bacterium (SRB) *Desulfovibrio vulgaris* Hildenborough (DvH). However, focusing on a single representative species of a genus offers little insight into the evolution, niche adaptation, and function of an entire microbial genus. Given the broad environmental significance of SRBs and their phylogenetic diversity, it is imperative that other representative species of the *Desulfovibrio* genus are investigated. Therefore, we are developing an integrated functional genomics platform in the SRB *Desulfovibrio desulfuricans* G20 (G20). The platform is compromised of phenotypic arrays, gene expression arrays, and a transposon mutant library. The ability of G20 to utilize carbon sources and electron acceptors for growth, survive stress conditions, and respond to growth inhibitors is measured on a high-throughput system to narrow the test conditions to be used in both gene expression and mutant profiling assays. Using high-density multiplex microarrays, we are monitoring gene expression under a number of diverse conditions in a global effort to elucidate G20 gene regulation. Our principle effort is the construction of a genome-wide, tagged transposon mutant library in G20. A hallmark of our transposon library is the ability to pool thousands of mutants for parallel phenotypic analysis. Profiling our transposon library under a large number of conditions will enable us to link a large fraction of the G20 genome to specific cellular phenotypes. Here we present our experimental approaches and preliminary data from each type of experiment.

Functional genomic investigations into G20 will be used by the ESPP in a number of ways. First, G20 and DvH have a large number of orthologous genes. Therefore, functional insights made in G20 are a strong starting point for targeted experiments in DvH. Second, by expanding our experimental "pan-genome" to include G20, we are now investigating more of the natural variation present in the genus *Desulfovibrio*. These data will be imperative for annotating and analyzing newly sequenced SRB genomes. Lastly, our integrated functional genomics data will be used to build a cellular model of G20 metabolism, regulation, and activity. This model can be used as a framework for the interpretation of many types of ESPP data generated in the laboratory and in the field.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

A Phylogenomic Approach to the Evolutionary Origins of Microbial Metabolisms

Lawrence David[1]* (ldavid@mit.edu) and Eric Alm[1,2]

[1]Massachusetts Institute of Technology, Cambridge, Mass.; and [2]The Virtual Institute for Microbial Stress and Survival, Berkeley, Calif.

Project Goals: This work is consistent with the DOE stated goal "to gain insights about fundamental biological processes."

We have developed a tool, AnGST, to infer the evolutionary history of gene families independent of species phylogenies. This allows for 'birth dates' to be assigned to individual genes. We use this genetically encoded history of functional genetic material to infer the types of molecular functions present on the Earth as a function of time. For example, enzymes utilizing molecular oxygen are found to be largely absent until ~2.7 Gya, consistent with geological estimates of the origins of oxygenic photosynthesis.

GTL

Resource for the Exploration of Regulons Accurately Predicted by the Methods of Comparative Genomics

Pavel S. Novichkov[1]* (PSNovichkov@lbl.gov), Mikhail S. Gelfand,[2] Inna Dubchak,[1,4] and Dmitry A. Rodionov[2,3]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; [3]Burnham Institute for Medical Research, La Jolla, Calif.; and [4]DOE Joint Genome Institute, Walnut Creek, Calif.

Project Goals: The current version of database covers more than 250 genomes and 180 profiles. Among others, it represents the results of our recent comparative genomic reconstruction of metabolic regulons in 13 *Shewanella* species that included near 70 transcription factors, approximately 400 binding sites and more than 1000 target genes per each genome. The database gives access to large regulatory networks reconstructed for certain metabolic pathways, e.g. degradation of fatty acids, branch chain amino acids, and aminosugars, homeostasis of biometals, and biosynthesis of NAD cofactor. In the near future we are planning to add a large collection of regulons for the LacI family transcription factors.

Identification and reconstruction of various transcriptional regulons in bacteria using a computational comparative genomics approach is coming of age. During the past decade a large number of manually-curated high quality inferences of transcriptional regulatory interactions were accumulated for diverse taxonomic groups of bacteria. These data provide a good foundation for understanding molecular mechanisms of transcriptional regulation, identification of regulatory circuits, and interconnections among circuits within the cell. Traditional experimental methods for regulon analysis have certain limitations both in terms of productivity and feasibility. While the development of high-throughput transcriptome approaches allow to obtain genome-scale gene expression patterns, in many cases the complexity of the interactions between regulons makes it difficult to distinguish between direct and indirect effects on transcription. The availability of a large number of closely related genomes allows one to apply comparative genomics to accurately expand already known regulons to yet uncharacterized organisms, and to predict and describe new regulons. Due to fast accumulation of such valuable data, there is a need for a specialized database and associated analysis tools that will compile and present the growing collection of high quality predicted bacterial regulons.

The RegPrecise database was developed for capturing, visualization and analysis of transcription factor regulons that were reconstructed by the comparative genomic approach. The primary object of the database is a single regulon in a particular genome, which is described by the identified transcription factor, its DNA binding site model (a profile), as well as the set of regulated genes, operons and associated operator sites. Regulons for orthologous transcription factors from closely related genomes are combined into the collections that provide an overview of the conserved and variable components of the regulon. A higher level representation of the regulatory interactions is also provided for orthologous regulons described in several bacterial taxonomic groups enabling comparison and evolutionary analysis of the transcription factor binding motifs. Another view of complex data in the database is a general overview of multiple regulons inferred in a set of closely related group of genomes.

The current version of database covers more than 250 genomes and 180 profiles. Among others, it represents the results of our recent comparative genomic reconstruction of metabolic regulons in 13 *Shewanella* species that included near 70 transcription factors, approximately 400 binding sites and more than 1000 target genes per each genome. The database gives access to large regulatory networks reconstructed for certain metabolic pathways, e.g. degradation of fatty acids, branch chain amino acids, and aminosugars, homeostasis of biometals, and biosynthesis of NAD cofactor. In the near future we are planning to add a large collection of regulons for the LacI family transcription factors.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Expression Profiling of Hypothetical Genes in *Desulfovibrio vulgaris* Leads to Improved Functional Annotation

Dwayne A. Elias* (eliasd@missouri.edu) Elliott C. Drury, Alyssa M. Redding, Aindrila Mukhopadhyay, Marcin Joachimiak, Huei-Che B.Yen, Matthew W. Fields,[3] Terry C. Hazen,[1] **Adam P. Arkin** (APArkin@lbl.gov), Jay D. Keasling,[1] and Judy D. Wall[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]University of Missouri, Columbia, Mo.; and [3]Montana State University, Bozeman, Mont.

Project Goals: Hypothetical and conserved hypothetical genes account for >30% of sequenced bacterial genomes. For the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough, 348 of the 3534 genes are annotated as conserved hypothetical (9.7%) with 889 hypothetical genes (25.0%). Given this large genome fraction, it is plausible that some genes serve critical cellular functions. The goals of this study were to determine which genes can be expressed and to provide a more functionally based annotation.

Hypothetical and conserved hypothetical genes account for >30% of sequenced bacterial genomes. For the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough, 348 of the 3534 genes are annotated as conserved hypothetical (9.7%) with 889 hypothetical genes (25.0%). Given this large genome fraction, it is plausible that some genes serve critical cellular functions. The goals of this study were to determine which genes can be expressed and to provide a more functionally based annotation. To accomplish this, expression profiles of the 1237 hypothetical and conserved hypothetical genes were obtained from transcriptomic datasets of 10 environmental stresses, complimented with iTRAQ proteomic data. Genes were divided into putatively polycistronic operons and those predicted to be monocistronic, then classified by basal expression levels and grouped according to changes in expression for one or multiple stresses. 1219 of these genes were transcribed with 265 proteins detected. There was no evidence for expression of 17 predicted genes. Except for the latter, annotation of all monocistronic genes was expanded using the above criteria and COG information. Polycistronic genes were annotated with expression information including proximity to more confidently annotated genes. Two targeted deletion mutants were used as test cases to determine the accuracy of the inferred functional annotations.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Impact of Elevated Nitrate on Sulfate-Reducing Bacteria: Implications of Inhibitory Mechanisms in Addition to Osmotic Stress

Qiang He,[1,2]* Zhili He,[3,13] Dominique C. Joyner,[4,13] Marcin Joachimiak,[5,13] Morgan N. Price,[5,13] Zamin K. Yang,[6,13] Huei-Che Bill Yen,[7,13] Christopher L. Hemme,[3,13] Romy Chakraborty,[4,13] Wenqiong Chen,[8,13] Matthew M. Fields,[9,13] David A. Stahl,[12,13] Jay D. Keasling,[5,11,13] Martin Keller,[6,13] **Adam P. Arkin**[5,10,13] (APArkin@lbl.gov), Terry C. Hazen,[4,13] Judy D. Wall,[7,13] and Jizhong Zhou[3,13]

[1]Dept. of Civil and Environmental Engineering and [2]Center for Environmental Biotechnology, University of Tennessee, Knoxville, Tenn.; [3]Institute for Environmental Genomics, Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.; [4]Earth Sciences Division and [5]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [6]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [7]Depts. of Biochemistry and Molecular Microbiology & Immunology, University of Missouri, Columbia, Mo.; [8]Diversa Corp, San Diego, Calif.; [9]Dept. of Microbiology, Montana State University, Bozeman, Mont.; [10]Dept. of Bioengineering and [11]Dept. of Chemical Engineering, University of California, Berkeley, Calif.; [12]Dept. of Civil and Environmental Engineering, University of Washington, Seattle, Wash.; and [13]Virtual Institute for Microbial Stress and Survival

Sulfate-reducing bacteria are studied for their potential in heavy metal bioremediation. However, the occurrence of elevated nitrate in contaminated environments has been shown to inhibit sulfate reduction activity. While the inhibition has been suggested to result from competition with nitrate-reducing bacteria, the possibility of direct inhibition of sulfate reducers by elevated nitrate needs to be explored. Using *Desulfovibrio vulgaris* as a model sulfate-reducing bacterium, it was observed that significant growth inhibition was effected by 70 mM $NaNO_3$ but not 70 mM NaCl, indicating the presence of inhibitory mechanisms in addition to osmotic stress. While the differential expression of a small number of genes in response to nitrate suggested the potential involvement of osmotic and nitrite stress responses, the roles of these two stress responses appear minor given the lack of similarity in the overall transcriptional profiles between nitrate, nitrite, and NaCl stress responses. The presence of unique stress response pathways in nitrate stress is further suggested by the lack of extensive similarities in the response profiles between nitrate stress and various other stress conditions. In addition, the importance of genes with functions in the metabolism of S-adenosylmethionine in the shift of energy flow was implicated in nitrate stress response.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through

contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

A Role of CO and a CO Sensor Protein in the Energy Metabolism of *D. vulgaris* Hildenborough

Lara Rajeev[1]* (lrajeev@u.washington.edu), Sergey Stolyar,[1] David A Stahl,[1] Grant Zane,[2] and Judy Wall[2]

[1]University of Washington, Seattle, Wash.; and [2]University of Missouri, Columbia, Mo.

Project Goals: We put forth the hypothesis that Coo and Hyn1 (NiFe) hydrogenases are involved in hydrogen production from lactate and pyruvate, and in the absence of either hydrogenase, electrons from lactate and pyruvate are directed towards CO production.

Recent studies suggest that carbon monoxide (CO) may play an important role in the energy metabolism of sulfate-reducing bacteria. The membrane bound cytoplasmically oriented CooMKLUXHF hydrogenase of *D. vulgaris* Hildenborough is similar to the Coo hydrogenases of *R. rubrum* and *C. hydrogenoformans*. In the latter systems, the Coo hydrogenases together with the CO dehydrogenase (CODH) oxidize CO to CO_2 and H_2. Both operons are regulated by a CO sensing transcriptional regulator CooA. The *D. vulgaris* Hildenborough genome also encodes genes for CODH and CooA. Predicted binding sites for CooA are located upstream of both CODH and the Coo hydrogenase operons. To determine if DvH CooA also acts as a CO sensor and if Coo hydrogenase is CO-regulated, we tested a mutant deleted for the *cooA* gene for growth on lactate-sulfate (LS) or pyruvate-sulfate (PS) in the presence of CO. With 1% CO in the headspace, wild type DvH grows efficiently and consumes CO, whereas a *cooA* mutant does not oxidize the CO. Interestingly, the *cooA* mutant grew efficiently on PS with CO in the headspace, but on LS the growth was poor and was inhibited by high H_2 accumulation. The *cooL* mutant lacking an active Coo hydrogenase was able to consume CO. Both wt DvH and the *cooL* mutant, but not the *cooA* mutant, were able to grow on CO as the sole energy source, although growth was very slow. The *cooA* mutant could grow syntrophically on lactate with *M. maripaludis*, whereas the *cooL* mutant could not grow. These observations suggest that the *coo* hydrogenase expression is not CO- or CooA-dependent, unlike that seen in the *R. rubrum* system.

To determine the physiological role of CO in DvH metabolism, we followed the fermentation burst in various DvH hydrogenase mutants. Wt DvH produces very little CO during growth. However, the *cooL* mutant shows a pronounced CO burst during growth on both LS and PS. Other mutants lacking either the Ech hydrogenase or the Hyd (Fe) hydrogenase showed no CO burst, but the *hyd hyn1* double mutant showed a CO burst. Notably, the *cooA* mutant did not accumulate CO during growth. We put forth the hypothesis that Coo and Hyn1 (NiFe) hydrogenases are involved in hydrogen production from lactate and pyruvate, and in the absence of either hydrogenase, electrons from lactate and pyruvate are directed towards CO production.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Comparison of the Sulfate-Reducing Capacity of *Desulfovibrio vulgaris* Hildenborough Deleted for the Operon Containing *qmoABC* and a Hypothetical Protein (DVU0851) versus Deletion of the Hypothetical Protein Alone

Grant M. Zane,* Dwayne A. Elias, Huei-Che Yen, and Judy D. Wall

Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov)

Project Goals: Comparison of the sulfate-reducing capacity of *Desulfovibrio vulgaris* Hildenborough deleted for the operon containing qmoABC and a hypothetical protein (DVU0851) versus deletion of the hypothetical protein alone.

Deletion of the operon encoding *qmoABC* and a hypothetical protein (HP) (DVU0848-51) in the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough resulted in an inability to respire sulfate. No suppressed mutants appeared in cultures of the deletion strain incubated in the presence of sulfate. Curiously, the Δ(*qmoABC* HP) mutant was also unable to ferment pyruvate. Respiration of sulfate and fermentation of pyruvate was restored by complementation with the *qmoABC*, HP genes. In order to determine the contribution of the promoter-distal hypothetical protein to the ability of this organism to reduce sulfate and ferment pyruvate, a second deletion was made of this gene alone. Although the mutant deleted for the single gene was able to reduce sulfate, it grew more slowly than wild-type and was stimulated by methionine and cysteine. Complementation restored growth to near wild-type levels.

This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

The Molecular Mechanism of Adaptation to Salt Stress Revealed by the Long-Term Evolution of *Desulfovibrio vugaris* Hildenborough

Aifen Zhou,[1,6]* Zhili He,[1,6] Marcin P. Joachimiak,[2,6] Paramvir S. Dehal,[2,6] **Adam P. Arkin**[2,6] (APArkin@lbl.gov), Kristina Hillesland,[4,6] David Stahl,[4,6] Judy Wall,[5,6] Terry C. Hazen,[3,6] and Jizhong Zhou[1,6]

[1]Institute for Environmental Genomics, Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.; [2]Physical Biosciences Division and [3]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [4]Civil & Environmental Engineering, University of Washington, Seattle, Wash.; [5]Depts. of Biochemistry and Molecular Microbiology & Immunology, University of Missouri, Columbia, Mo.; and [6]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Project Goals: In order to investigate the long-term evolutionary responses, diversifications and adaptation to salt stress, sulfate reducing bacteria *Desulfovibrio vugaris* Hildenborough, a model environmental organism, is used to address such fundamental questions by mimicking the stress condition in the lab culture.

One of the greatest challenges in biology is to understand the interaction between genotype and environment to determine the fitness of an organism. With the recent advances in genome sequencing and high-throughput genomic technologies, now it is possible to link sub-cellular molecular/metabolic processes with the population-level processes, functions and evolution. In order to investigate the long-term evolutionary responses, diversifications and adaptation to salt stress, sulfate reducing bacteria *Desulfovibrio vugaris* Hildenborough, a model environmental organism, is used to address such fundamental questions by mimicking the stress condition in the lab culture. Control lines and stressed lines (6 lines each, from single colony based pure culture) grown in medium LS4D and LS4D + 100 mM NaCl were transferred every 48 hrs with one to one hundred dilutions. Phenotype of all the cell lines in terms of salt tolerance was tested with LS4D supplemented with 250 mM NaCl. Results demonstrated that the adaptation to salt stress is a dynamical process. The enhanced salt tolerance of stressed lines was observed at 300 generations and became more obvious with the increase of generations. The de-adaptation experiment on 500, 1000 and 1200 generation cell lines not only provided strong evidence that the phenotype was due to the genetic change instead of physiological adaptation, but also indicated that there is also a dynamic trend for genetic adaptation and the genetic mutation might become stable at 1000 generation. In order to further understand the molecular mechanisms of adaptation to salt stress in long-term evolution process, gene expression profiles of the 500 and 1000 generation samples were examined by *D. vulagris* whole genome oligo microarray. "Energy production and conversion" and "signal transduction mechanisms" are among the gene categories with most genes up-regulated. Statistical analysis also showed that gene expression profiling between evolved lines with salt stress vs control evolved lines and the evolved lines with salt stress vs ancestor are more closer at 1000 generation. Whole genome sequencing on selected colonies is underway to identify the beneficial genetic mutation.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Desulfovibrio vulgaris Hildenborough Responses to Salt and H_2O_2 Stresses

Zhili He,[1,10]* Aifen Zhou,[1,10] Qiang He,[2,10] Aindrila Mukhopadhyay,[3,10] Edward Baidoo,[3,10] Marcin Joachimiak,[3,10] Christopher L. Hemme,[1,10] Peter Benke,[3,10] Alyssa M. Redding,[3,10] Matthew M. Fields,[4,11] David A. Stahl,[5,10] Jay D. Keasling,[3,6,10] **Adam P. Arkin**[3,7,10] (APArkin@lbl.gov), Terry C. Hazen,[8,10] Judy D. Wall,[9,10] and Jizhong Zhou[1,10]

[1]Institute for Environmental Genomics, Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.; [2]Dept. of Civil and Environmental Engineering, University of Tennessee, Knoxville, Tenn.; [3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [4]Dept. of Microbiology, Montana State University, Bozeman, Mont.; [5]Dept. of Civil and Environmental Engineering, University of Washington, Seattle, Wash.; [6]Dept. of Chemical Engineering, University of California, Berkeley, Calif.; [7]Dept. of Bioengineering, University of California, Berkeley, Calif.; [8]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [9]Depts. of Biochemistry and Molecular Microbiology & Immunology, University of Missouri, Columbia, Mo.; and [10]Virtual Institute for Microbial Stress and Survival

Project Goals: Diverse stress resistance mechanisms may be used in *D. vulgaris* for detoxification of H_2O_2 with the up-regulation of DNA repair systems and the down-regulation of energy metabolism and protein synthesis.

The response of *Desulfovibrio vulgaris* Hildenborough to salt and H_2O_2 stresses were examined by physiological, global transcriptional, metabolite, and mutagenesis analyses. The growth of *D. vulgaris* was inhibited by 250 mM NaCl or 1 mM H_2O_2. Salt adaptation (long-term NaCl exposure) increased the expression of genes involved in amino acid biosynthesis and transport, electron transfer, hydrogen

oxidation, and general stress responses (e.g., heat shock proteins, phage shock proteins, and oxidative stress response proteins). Genes involved in carbon metabolism, cell motility, and phage structures were decreased in expression. Comparison of transcriptomic profiles of *D. vulgaris* responses to salt adaptation with those of salt shock (short-term NaCl exposure) showed some similarity as well as a significant difference. Metabolite assays showed that glutamate and alanine accumulated under salt adaptation, suggesting that they may be used as osmoprotectants in *D. vulgaris*. Addition of amino acids (glutamate, alanine, tryptophan) or yeast extract to the growth medium relieved salt-related growth inhibition. A conceptual model is proposed to link the observed results to currently available knowledge for further understanding the mechanisms of *D. vulgaris* adaptation to elevated NaCl. Under H_2O_2 conditions, PerR regulon genes were significantly up-regulated, indicating the importance role of PerR in oxidative stress response. In addition, some Fur regulon genes were also strongly induced. Increased gene expression of thiol-peroxidase genes ahpC as well as thioredoxin reductase and thioredoxin genes indicated the involvement of thiol switch in the oxidative stress response. rbr2 was the only significantly up-regulated H_2O_2 scavenging enzymes. The oxidative stress response of mutants ΔperR and Δfur demonstrated that ahpC and rbr2 were regulated by both Fur and PerR. The links between the up-regulated genes involved in H_2O_2 scavenging, protein fate, DNA metabolism and lipid metabolism and the down-regulated genes involved in sulfate reduction, energy production and translation were demonstrated by the gene co-expression network. The proteomics data provided further evidence at the translational level and complemented the transcriptomics data. Taken together, diverse stress resistance mechanisms may be used in *D. vulgaris* for detoxification of H_2O_2 with the up-regulation of DNA repair systems and the down-regulation of energy metabolism and protein synthesis.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Analysis of an Intact Dissimilatory Sulfite Reductase Protein Complex from *Desulfovibrio vulgaris* Using an Ion Mobility QTOF Analyzer

Ming Dong,[1] Michael Daly,[2] Haichuan Liu,[3] Simon Allen,[3] Evelin Szakal,[3] Steven C. Hall,[3] Susan J. Fisher,[1,3] Terry C. Hazen,[1] Jil T. Geller,[1] Mary E. Singer,[1] Lee L. Yang,[1] Jian Jin,[1] H. Ewa Witkowska,[2]* and **Mark D. Biggin**[1] (mdbiggin@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Waters Corporation; and [3]University of California, San Francisco, Calif.

Project Goals: MS analysis of intact and denatured Dsr complex was performed utilizing HD Synapt mass spectrometer. Three forms of the complex were isolated from a 400 L DvH culture using a 3-step protein fractionation protocol. MS analysis demonstrated that three forms of Dsr complexes differed in stoichiometry: the presence of a hexamer $[DsrA]_2[DsrB]_2[DsrC]_2$, a pentamer $[DsrA]_2[DsvB]_2[DsvC]$ and a tetramer $[DsrA]_2[DsvB]_2$, all measured within an error of less than 0.1%, was demonstrated. No differences in molecular masses of subunits participating in three distinct forms of Dsr complex were detected. However, all three subunits demonstrated discrepancies between the experimental and theoretical molecular masses. Analysis of DsrC subunit revealed the presence of post-translational modifications.

Protein extract was isolated from a 400 L *D. vulgaris* culture grown under optimal conditions. Dsr complexes were purified utilizing three sequential steps of fractionation: ammonium sulfate precipitation, anion exchange chromatography (Mono Q) and size exclusion chromatography (Superdex 200). The purity of the preparation was 95%, based on SDS PAGE. Intact complexes in 30 mM ammonium bicarbonate were infused by nanospray into a Synapt HD MS instrument that combines ion mobility, quadrupole and TOF analyzers. In addition, Dsr complexes were analyzed under denaturing conditions utilizing on-line LC MS and MS/MS on the same instrument.

Dsr was detected in three distinct Mono Q column fractions: two of the early-eluting fractions contained all three subunits (DsrA, DsrB and DsrC) while DsrC polypeptide was absent in the late-eluting fraction. No differences in molecular masses of polypeptide constituents participating in the three different complex forms were revealed by ESI MS, within method error approximating 10 ppm. However, all three polypeptides, DsrA, DsrB and DsrC, demonstrated a significant discrepancy between the theoretical and experimental molecular masses suggesting the presence of post translational modifications (PTMs) and/or sequence errors. To explain this discrepancy, DsrC was analyzed in detail. Gas phase fragmentations in the source produced top-down sequence information that demonstrated that the

regions [1-36] and [84-104] were consistent with the DNA sequence. In addition, the presence of a disulfide bridge between Cys 92 and Cys 103 was revealed. Thus the above results pointed to the region [37-83] as a site of putative PTMs. The presence of either trimethylation or acetylation on two Lys residues was demonstrated by LC MS/MS analysis of tryptic peptides derived from DsrC via in-gel digestion, also reported by Gaucher et al.(4).

Molecular mass measurements of intact complexes demonstrated a difference in stoichiometry between the two early eluting dissimilatory sulfite reductase forms: the first fraction contained a hexamer $[DsrA]_2[DsrB]_2$ $[DsrC]_2$ while the second fraction contained a pentamer $[DsrA]_2$ $[DsvB]_2$ [DsvC], both measured within an error of less than 0.1%. The third fraction contained a tetramer $[DsrA]_2$ $[DsvB]_2$. The accuracy of mass measurement allowed for assignment of the number of polypeptide subunits within each complex. However, the exact number and structure of the prosthetic groups involved was not discerned. The products of partial protein complex dissociation (either in solution or in the gas phase) suggested that the [DsrA][DsrB] carrying prosthetic groups formed the core of all three forms of Dsr complex.

References

1. Crane and Getzoff (1996) *Curr. Op. Struct. Biol.* 6:744
2. Pierik et al, (1992) *Eur. J. Biochem.* 205:11
3. Marritt and Hagen (1996) *Eur J Biochem*, 238:724
4. Gaucher et al. (2008) *J. Proteome Res.* 7: 2320

GTL

Microbes Online: An Integrated Portal for Comparative Functional Genomics

Paramvir S. Dehal[1,2]* (PSDehal@lbl.gov), Jason K. Baumohl,[1,2] Dylan Chivian,[1,2,3] Katherine H. Huang,[1,2] Marcin P. Joachimiak,[1,2] Keith Keller,[1,2] Morgan N. Price,[1,2] and **Adam P. Arkin**[1,2,3,4]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov; [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [3]DOE Joint BioEnergy Institute, Emeryville, Calif.; and [4]Dept. of Bioengineering, University of California, Berkeley, Calif.

Project Goals: MicrobesOnline seeks to integrate functional genomic data with comparative genome sequence to provide novel web-based viewing and analysis tools for gene expression microarray, proteomic, and phenotype microarray data. Selecting an organism or gene of interest in MicrobesOnline leads to information about and data viewers for experiments conducted on that organism and involving that gene or gene product. Among the major new features is the ability to search the microarray data compendium for genes with gene expression profiles similar to a query expression profile (either based on a gene or set of genes).

Since 2003, MicrobesOnline (http://www.microbesonline.org) has been providing a community resource for comparative and functional genome analysis. The portal includes over 1000 complete genomes of bacteria, archaea and fungi, as well as 1000s of viruses and plasmids. In addition to standard comparative genomic analysis, including gene prediction, sequence homology, domain identification, gene family assignments and functional annotations from E.C. and GO, MicrobesOnline integrates data from functional genomics and places it within a phylogenetic context. Currently, MicrobesOnline has 1000s of microarray based expression experiments from diverse set of organisms ranging from model organisms such as *E. coli* and *S. cerevisiae* to environmental microbes such as *Desulfovibrio vulgaris* and *Shewanella oneidensis*. MicrobesOnline offers a suite of analysis and tools including: a multi-species genome browser, operon and regulon prediction methods and results, a combined gene and species phylogeny browser, phylogenetic profile searches, a gene expression data browser with gene expression profile searches, a gene ontology browser, a workbench for sequence analysis (including sequence motif detection, motif searches, sequence alignment and phylogeny reconstruction), integration with RegTransBase, and capabilities for community annotation of genomes. The next update of MicrobesOnline will contain significant new functionality, including comparative analysis of metagenomic sequence data.

MicrobesOnline seeks to integrate functional genomic data with comparative genome sequence to provide novel web-based viewing and analysis tools for gene expression microarray, proteomic, and phenotype microarray data. Selecting an organism or gene of interest in MicrobesOnline leads to information about and data viewers for experiments conducted on that organism and involving that gene or gene product. It is possible to view microarray data from multiple conditions as an interactive heatmap and to analyze correlations between gene expression results from different experiments. Among the major new features is the ability to search the microarray data compendium for genes with gene expression profiles similar to a query expression profile (either based on a gene or set of genes). These new compendium-wide functionalities allow the user to observe patterns in gene expression changes across multiple conditions and genes, and to search for similarities to these patterns. The information integration and analysis performed by MicrobesOnline serves not only to generate insights into the gene expression responses and their regulation in these microorganisms, but also to document experiments, allow contextual access to experimental data, and facilitate the planning of future experiments. MicrobesOnline is actively incorporating publicly available functional genomics data from published research, so as to centralize data on microbial physiology and ecology in a unified comparative functional genomic framework.

The Virtual Institute for Microbial Stress and Survival (VIMSS, http://vimss.lbl.gov) funded by the Dept. of Energy's Genomics:GTL Program, is dedicated to using integrated environmental, functional genomic, comparative sequence and phylogenetic data to understand mechanisms

by which microbes survive in uncertain environments while carrying out processes of interest for bioremediation and energy generation. To support this work, VIMSS has developed this web portal (MicrobesOnline), an underlying database, and analyses for comparative functional genomics of bacteria, archaea, fungi and viruses.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Progress in the Development of the RegTransBase Database and the Comparative Analysis System

Michael J. Cipriano,[1] Pavel Novichkov,[1,4] Alexei E. Kazakov,[2] Dmitry Ravcheev,[2] **Adam P. Arkin**[1,3,4] (APArkin@lbl.gov), Mikhail S. Gelfand,[2,5] and Inna Dubchak[1,4,6*] (ILDubchak@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Institute for Information Transmission Problems, RAS, Moscow, Russia; [3]Dept. of Bioengineering, University of California, Berkeley, Calif.; [4]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov, Berkeley, Calif.; [5]Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; and [6]DOE Joint Genome Institute, Walnut Creek, Calif.

RegTransBase, a database describing regulatory interactions in prokaryotes, has been developed as a component of the MicrobesOnline/RegTransBase framework successfully used for interpretation of microbial stress response and metal reduction pathways. It is manually curated and based on published scientific literature. RegTransBase describes a large number of regulatory interactions and contains experimental data which investigates regulation with known elements. It is available at http://regtransbase.lbl.gov.

Over 1000 additional articles were annotated last year resulting in the total number of 5118 articles. We specifically focused on annotating the facts of regulation in bioenergy-related bacteria such as: *Clostridia*, *Thermoanaerobacter*, *Geobacillus stearothermophilus*, *Zymomonas*, *Fibrobacter*, *Ruminococcus*, *Prevotella*, *Acetobacter*, *Anaeromyxobacterium*, *Streptomyces*, *Ralstonia*.

Currently, the database describes close to 12000 experiments (30% growth in the last year) in relation to 531 genomes. It contains data on the regulation of ~39000 genes and evidence for ~10000 interactions with ~1130 regulators. We removed redundancy in the list of Effectors (currently the database contains about 630 of them) and turned them into controlled vocabulary.

RegTransBase additionally provides an expertly curated library of 150 alignments of known transcription factor binding sites covering a wide range of bacterial species. Each alignment contains information as to the transcription factor which binds the DNA sequence, the exact location of the binding site on a published genome, and links to published articles. RegTransBase builds upon these alignments by providing a set of computational modules for the comparative analysis of regulons among related organisms.

The new tool—"advanced browsing" was developed to allow a user to search the data contained in RegTrandBase in a step-by-step manner. Different types of classifications, such as taxonomy, effectors, the type of experimental result, a phenotype, and genome relevance allow for creating and applying complex search criteria. We are planning to include additional classifications such as metabolic pathways and types of experimental techniques in this scheme.

There is an increasingly tight coupling of RegTransBase with MicrobesOnline in reporting cis-regulatory sites and regulatory interactions, and integrating RegTransBase searches into MicrobesOnLine cart functions.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Applied Environmental Microbiology Core Research on Stress Response Pathways in Metal-Reducers VIMSS: ESPP

Terry C. Hazen[1,2*] (TCHazen@lbl.gov), Gary Anderson,[1,2] Sharon Borglin,[1,2] Eoin Brodie,[1,2] Steve van Dien,[8] Matthew Fields,[1,7] Julian Fortney,[1,2] Jil Geller,[1,2] E. Hendrickson,[5] Kristina L Hillesland,[1,6] Hoi-Ying Holman,[1,2] J. Leigh,[1,6] T. Lie,[1,6] Janet Jacobsen,[1,2] Dominique Joyner,[1,2] Romy Chakraborty,[1,2] Martin Keller,[1,3] Aindrila Mukhopadhyay,[1,2] Christopher Schadt,[1,3] David Stahl,[1,6] Sergey Stolyar,[1,6] Chris Walker,[1,6] Judy Wall,[1,5] Zamin Yang,[1,3] Huei-che Yen,[1,5] Grant Zane,[1,5] and Jizhong Zhou[1,9]

[1]Virtual Institute of Microbial Stress and Survival; [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [4]Diversa, Inc., San Diego, Calif.; [5]University of Missouri, Columbia, Mo.; [6]University of Washington, Seattle, Wash.; [7]Montana State University, Bozeman, Mont.; [8]Genomatica, San Diego, Calif.; and [9]University of Oklahoma, Norman, Okla.

Project Goals: The goals of this project are: (i) to develop a stable syntrophic microbial consortia in continuous

flow systems which can be used for physiological and functional genomic studies consisting of *Clostridium cellulolyticum*, *Desulfovibrio vulgaris*, *Geobacter sulfurreducens* and *Methanococcus maripaludis* to simulate community interactions in anaerobic systems, (ii) to study how organisms change their activity on a molecular level in response to other community members in the co-culture by analyzing the gene transcription, metabolite flux and growth, and (iii) to study how these changes in activity and community organization affect the resistance and resilience of microbial communities in response to stress, invasion and other perturbations on a molecular level.

Field Studies

Environmental Characterizations. Previous research specifically points toward SRB as environmentally relevant experimental systems for the study of heavy metal and radionuclide reduction, and our recent data has detected *Desulfovibrio* sequences at the FRC and Hanford 100H. To effectively immobilize heavy metals and radionuclides, it is important to understand the cellular responses to adverse factors observed at contaminated subsurface environments, such as mixed contaminants and the changing ratios of electron donors and acceptors. In a recent study, we focused on responses to Cr(VI). At Hanford 100H as part of our ERSP project we injected 40 lbs of HRC (polylactate) as a slow release electron donor in August 2004. Until March 2008 reducing conditions were maintained, along with undetectable levels of Cr(VI) (Hubbard et al, 2008; Faybishenko et al., 2008). During this time the environment was dominated by sulfate reducers and we were able to detect *Desulfovibrio vulgaris*-like organisms with direct fluorescent antibody. We also isolated a strain of *D. vulgaris* that was only 100 genes different from the type strain (this strain is currently being sequenced by JGI). This year we injected 10 lbs of HRC at the same site to determine if there is a 'memory' response and observed H_2S production after only 23 h. Once the community stabilizes we will begin push-pull stress tests in the field with NO_3 and monitor functional gene, community structure, and stress responses as compared with previously published models by our group with pure cultures. We are also isolating consortia and determining the dominant community structure to compare with our lab studies. To determine if the stress response in biofilm communities is different then groundwater we are also using "bug traps". Bug traps were filled with Hanford sediment in order to stimulate surface-associated populations in the 100-H subsurface. Work is in progress, and initial results suggest that even under non-stimulated conditions, that more nucleic acids can be extracted from sediment compared to an equivalent amount of groundwater.

Technique Development for Environmental DNA and mRNA analysis. This year we did further optimization of the MDA approach to isolate and amplify DNA form samples with extremely low biomass. We used Hanford 100H samples to construct environmental libraries for sequencing and screening. We also evaluated three different methods in pure cultures to remove rRNA and tRNA from samples in order to screen mRNA expression that will eventually be applied to environmental samples. The first method utilizes biotin-modified oligos complementary to conserved regions in 16S & 23S rRNA and subtractive hybridization with streptavidin-coated magnetic beads. The second uses a commercially available exonulclease that specifically digests rRNAs bearing a 5'monophosphate group. The third method uses two rounds of reverse transcription, where rRNAs are first reverse transcribed with multiple universal primers for 16S & 23S RNAs, subsequently the RNA/DNA hybrids and cDNA are removed by sequential digestion with RNaseH and DNaseI, and the enriched mRNAs are then reverse transcribed using random primers. We evaluated these three methods by comparing disappearance of the 16S and 23S bands via electrophoresis, and their effect on mRNA quality and quantity by analysis of transcription levels of control (total RNA) vs. enriched mRNA as measured whole genome microarray. Enriched mRNAs from the first two methods generated more genes with altered transcript levels compared to untreated total RNA, with 19 genes (0.5%) for the exonuclease method & 74 genes (2%) for subtractive hybridization exhibiting significant differences ($P < 0.05$).

Genome Sequence. The genome sequence for *Desulfovibrio vulgaris* strain DePue has been closed and annotated in collaboration with other ESPP investigators. This genome appears highly conserved and syntenous compared with *D. vulgaris* Hildenborough except for dramatic differences in bacteriophage content. Additionally, strain DePue contains a large (~50 kb) unique region encoding for an exopollysaccahride production, modification and transport system. This year we also submitted *D. vulgaris* Miyazaki, *D. salixigens*, *D. desulfuricans* 27774, *D. aesopensis*, *D. vulgaris* RCH1, *D. fructosivorans*, *D. hanfordensis*, *D.* sp. FW1012B, and *Sulfurospirillum barnesii* strain SES-3 for sequencing at JGI. We are also preparing *D. termitidis* for sequencing at JGI. *Desulfovibrio* FW1012B was isolated from ORFRC groundwater during biostimulation for uranium bioreduction. The isolate can reduce sulfate with lactate, pyruvate, and 1,2-propandiol, reduces chromate, and reduces nitrate without growth. In addition, the genome of *Anaeromyxobacter* fw109-5 has been completed. The *Anaeromyxobacter* strain was isolated from ORFRC sediments, and can reduce iron and heavy metals.

Pyro-sequencing. Methods are being developed for parsing environmental pyro-sequencing results for analyses with multivariate techniques. A test set was constructed of 3 samples that represented almost 19,000 SSU rDNA gene fragments.

Multispecies Syntrophic Consortia. The goals of this project are: (i) to develop a stable syntrophic microbial consortia in continuous flow systems which can be used for physiological and functional genomic studies consisting of *Clostridium cellulolyticum, Desulfovibrio vulgaris, Geobacter sulfurreducens* and *Methanococcus maripaludis* to simulate community interactions in anaerobic systems, (ii) to study how organisms change their activity on a molecular level in response to other community members in the co-culture by analyzing the gene transcription, metabolite flux and growth, and (iii) to study how these changes in activity and community organization affect the resistance and resilience

of microbial communities in response to stress, invasion and other perturbations on a molecular level. Stable community configurations and basic interaction structures within the community have recently been demonstrated via QPCR analysis of populations and HPLC/GCMS analysis of the major metabolite flux in the system. A four species microarray configuration is currently being evaluated and used for characterization of stable growth interactions transcriptional patterns in the consortia and comparison to the individual species grown in isolation. We have also found a role for carbon monoxide in the energy metabolism of *Desulfovibrio*. Combined physiological and genetic studies have shown a capacity for both production and consumption of carbon monoxide, and identified a CO sensor (CooA) that is required for CO consumption, most likely via CO-dependent regulation of a carbon monoxide dehydrogenase (CODH). We have also observed a rapid evolutionary improvement in the stability, growth rate, and production (biomass yield) of a simple model community comprised of *D. vulgaris* and *M. maripaludis* after 300 generations of evolution in either a uniform (shaken) or heterogeneous (static) environment. A contribution of both populations to evolutionary improvement was demonstrated. However, the relative contribution of each population to improvement was dependent on growth environment (uniform versus heterogeneous).

Acknowledgements
This work was part of the Virtual Ins4tute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Microfluidic Tools for Single-Cell Genomic Analysis of Environmental Bacteria

Robert J. Meagher,[1,3] Masood Z. Hadi,[1,3] Yooli K. Light,[1,3] Paramvir Dehal,[2,3] Terry Hazen,[2,3] **Adam Arkin**[2,3] (aparkin@lbl.gov), and Anup K. Singh[1,3]* (aksingh@sandia.gov)

[1]Sandia National Laboratories, Livermore, Calif.; [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [3]Virtual Institute for Microbial Stress and Survival, Berkeley, Calif.

Project Goals: 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. Three microfluidic technologies have been developed to enable these studies: (1) automated 16S rRNA FISH for identification of microbial cells in a mixed sample, (2) photonic force cell sorting for selectively isolating a species of interest (3) encapsulation of individual cells in picoliter-volume droplets, followed by genetic analysis.

Currently available experimental tools such as microarrays and qRT-PCR for studying the genomes and gene expression in mixed populations of environmental bacteria generally require relatively large amounts of starting material (DNA or RNA), and provide population-averaged data. Although these experimental tools provide valuable insights into microbial communities, the pooling of samples severs the link between the genotype and phenotype of each individual cell. 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. Three microfluidic technologies have been developed to enable these studies: (1) automated 16S rRNA FISH for identification of microbial cells in a mixed sample, (2) photonic force cell sorting for selectively isolating a species of interest (3) encapsulation of individual cells in picoliter-volume droplets, followed by genetic analysis. These three technologies can be coupled to one another, allowing, *e.g.* FISH-based identification of a rare species, followed enrichment of the rare species of interest by photonic force cell sorted, followed by single-cell encapsulation and PCR analysis. The droplet technology in particular allows us to scale down conventional (microliter-volume) assays, such as PCR, 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. Droplets can be generated rapidly (hundreds per second), with very good uniformity in size (<5% variation in droplet diameter), allowing high throughput experiments to be conducted with much better precision than is possible with conventional emulsion techniques. These technologies are currently under development with simple laboratory strains of *E. coli* and other well-characterized organisms such as *D. vulgaris*, and upon validation will then be translated to studying more complex mixed cultures and environmental samples.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Applications of GeoChip for Analysis of Different Microbial Communities

Joy D. Van Nostrand,[1,2]* Liyou Wu,[1] Patricia Waldron,[1] Ye Deng,[1] Zhili He,[1,2] Weimin Wu,[3] Sue Carroll,[4] Chris Schadt,[4,2] Anthony Palumbo,[4] Dave Watson,[4] Craig Criddle,[3] Phil Jardine,[4] Terry C. Hazen,[5,2] and Jizhong Zhou[1,2]

[1]University of Oklahoma, Norman, Okla.; [2]VIMSS (Virtual Institute for Microbial Stress and Survival), http://vimss.lbl.gov; [3]Stanford University, Stanford, Calif.; [4]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [5]Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: Microarray technology provides the opportunity to identify thousands of microbial genes or populations simultaneously. Based on GeoChip 2.0, a new version of GeoChip (GeoChip 3.0) has been developed, which has several new features. First, GeoChip 3.0 contains ~25,000 probes and covers ~47,000 sequences for 292 gene families. Second, the homology of automatically retrieved sequences by key words is verified by HUMMER using seed sequences so that unrelated sequences are removed. Third, a universal standard has been implemented so that data normalization and comparison of different microbial communities can be conducted. Fourth, a genomic standard is used to quantitatively analyze gene abundance.

Microarray technology provides the opportunity to identify thousands of microbial genes or populations simultaneously. Based on GeoChip 2.0, a new version of GeoChip (GeoChip 3.0) has been developed, which has several new features. First, GeoChip 3.0 contains ~25,000 probes and covers ~47,000 sequences for 292 gene families. Second, the homology of automatically retrieved sequences by key words is verified by HUMMER using seed sequences so that unrelated sequences are removed. Third, a universal standard has been implemented so that data normalization and comparison of different microbial communities can be conducted. Fourth, a genomic standard is used to quantitatively analyze gene abundance. In addition, GeoChip 3.0 includes phylogenic markers, such as gyrB. Finally, a software package has been developed to facilitate management of probe design, data analysis, and future updates. The GeoChip has been used to examine dynamic functional and structural changes in microbial communities from many different environments. Here, three examples of studies utilizing the GeoChip to examine microbial communities at contaminated sites are presented. These three studies examined areas within the U.S. DOE's Field Research Center (FRC) in Oak Ridge, TN. (1) Microbial communities within a pilot-scale test system established for the biostimulation of U (VI) reduction in the subsurface by injection of ethanol were examined. Sediments from eleven different sampling wells were evaluated. The results showed that different microbial communities were established in different wells and a high gene overlap was observed from wells within the same treatment zone. Higher microbial functional gene number, diversity and abundance were observed within the active bioremediation zone. The microbial community structure was highly correlated with the hydraulic flow rate and geochemical conditions of the treatment zone, especially pH, manganese concentration and electron donor level. (2) In a different study of the same system, functional community dynamics were examined during a period of oxidation by nitrate. Diversity and gene number decreased after nitrate exposure and while recovery appeared to begin, the gene numbers were still low even 100 days after nitrate exposure. Principle component analysis (PCA) of detected genes indicated a shift in community structure after nitrate exposure. Nitrate exposure appeared to result in long-term changes to the overall community. (3) In the third study from the FRC, analysis of groundwater monitoring wells along a contamination gradient revealed less overlap between wells with different levels of U and NO_3^- contamination. While diversity of nitrogen fixation genes decreased in NO_3^--contaminated wells, the diversity of metal reduction and resistance genes did not correlate with metal concentrations. Signal intensity did, however, increase in heavily contaminated wells, indicating a larger percentage of organisms with metal- related genes. Sulfate-reduction genes had greater diversity and greater signal intensity in more contaminated wells. Individual principle component analyses (PCA) of the gene diversity and geochemistry of the wells separated them in similar ways. CCA indicated that pH was an important variable that correlated with gene diversity in the lowest-contamination well, while NO_3^- and U correlated with the most highly contaminated well. Overall, contaminant level appears to have significant effects on the functional gene diversity along the contaminant plume at the FRC. These studies demonstrate the analytical power of the GeoChip in examining microbial communities. This is the first comprehensive microarray available for studying the functional and biogeochemical cycling potential of microbial communities, and it is also a powerful tool to link microbial community structures to ecosystem functioning.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

* Presenting author

GTL

Temporal and Spatial Organization within a Syntrophic Bacterial-Archaeal Biofilm

K. Brileya,[1] C. Walker,[2] S. Stolyar,[2] D.A. Stahl,[2] **A. Arkin**[3] (AParkin@lbl.gov), T.C. Hazen,[3] and M.W. Fields[1]* (matthew.fields@erc.montana.edu)

[1]Montana State University, Bozeman, Mont.; [2]University of Washington, Seattle, Wash.; and [3]Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: A goal of VIMSS is to determine the molecular determinants that underlie microbial community function and stability.

The elucidation of how populations interact in a given community and how the community responds to stress and perturbations can help infer the interplay between stress pathways and gene networks that help optimize bacterial biochemistry. A goal of VIMSS is to determine the molecular determinants that underlie microbial community function and stability. A syntrophic co-culture of the sulfate-reducing bacterium, *Desulfovibrio vulgaris* Hildenborough, and the methanogenic archaeon, *Methanococcus maripaludis*, was selected as a basal community that can directly and indirectly interact as a biofilm. Planktonic growth conditions, in which cells exist as 'non-adhered cells', rarely represent a true state of growth for the majority of microorganisms under *in situ* conditions, and adherent growth is most likely a universal feature. The roles of biofilms have become increasingly more evident in processes from microbial pathogenesis to waste water to metal corrosion; however, relatively little work has been done on anaerobic biofilms, particularly regarding the structure and behavior of non-pathogenic organisms under environmentally relevant conditions. Microbial communities associated with surfaces may incur protection from stresses such as nutrient-limitation, pH, salts, and heavy metals. In addition, proximity and localization within surface-adhered communities may impact functionality in terms of electron- and hydrogen-metabolism. It was hypothesized that hydrogen transfer would dictate co-culture biofilm formation in the absence of sulfate as terminal electron acceptor for *D. vulgaris* and without addition of hydrogen as electron donor for the methanogen. *M. maripaludis* did not form significant biofilms on a glass surface in batch mono-culture experiments, but *D. vulgaris* did. However, *M. maripaludis* did form a pellicle-like structure in batch, static cultures. A biofilm reactor was developed to co-culture *D. vulgaris* and *M. maripaludis* during syntrophic growth, and spatial and temporal organization was characterized using qPCR, epifluorescent microscopy, field emission electron microscopy, methane production and protein and carbohydrate analysis. During early development, the biofilm initiated as a monolayer of *D. vulgaris* cells, and the mainly *D. vulgaris* biofilm contained extracellular filaments that have been previously described. Soon after the development of the *D. vulgaris* biofilm, *M. maripaludis* cells were observed, and the number of planktonic phase cells declined as the number of biofilm cells increased for both populations. Over time, the methanogenic biofilm stabilized, and the ratio of *D. vulgaris* to *M. maripaludis* cells was approximately 2.5 and this is a similar ratio observed for cultures entirely populated by planktonic cells. However, at later time points, the planktonic populations had a ratio of approximately 0.2, and this ratio was significantly lower compared to biofilm. Both populations had 1 to 2 log more cells in the biofilm than the planktonic phase. As the methanogenic biofilm developed, extracellular structures continued to be observed. The results suggested that *D. vulgaris* initiated and established a biofilm that then recruited *M. maripaludis*, and the biofilm grew and changed over time as the numbers of both populations increased.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Evolution and Stability in a Syntrophic Community

Kristina L Hillesland* (hilleskl@u.washington.edu) and David A Stahl

University of Washington, Seattle, Wash.

Project Goals: We compared 20 independently evolved cocultures to cocultures with ancestral *D. vulgaris* and *M. maripaludis*. All but one of the evolved cocultures could grow significantly faster and obtain a much higher yield from the same resource. We tested whether these changes could be attributed to *D. vulgaris*, *M. maripaludis*, or both species by analyzing mixtures of evolved and ancestral populations and found that increases in coculture yield were caused by the combined adaptations of both species.

Mutualistic relationships between species may often originate from an opportunistic exchange of byproducts produced by each species, but few have addressed how microorganisms evolve in response to these interactions. We used experimental evolution to observe this process in real time in an interaction between the bacterium *Desulfovibrio vulgaris* and the archeaon, *Methanococcus maripaludis*. This mutualism based on the exchange of hydrogen represents an interaction that is fundamental to many microbial communities. After 300 generations of evolution, we compared 20 independently evolved cocultures to cocultures with ancestral *D. vulgaris* and *M. maripaludis*. All but one of the evolved cocultures could grow significantly faster and obtain a much higher yield from the same resource. We tested whether these changes could be attributed to *D. vulgaris*, *M. maripaludis*, or both species by analyzing mixtures of evolved and ancestral populations and found that increases in coculture yield were caused by the combined adaptations of both species. Whether *D. vulgaris*, *M. maripaludis*, or

both contributed to improvements in coculture growth rate depended on the heterogeneity of the evolution environment. Several cocultures had evolved in tubes that were not shaken, so that gradients of hydrogen could form during incubation. Both *D. vulgaris* and *M. maripaludis* that evolved in this environment were capable of enhancing syntrophic growth rate when paired with the ancestor, but *M. maripaludis* that evolved in a constantly shaken homogeneous environment had little if any effect. The evolutionary process leading to these adaptations involved dramatic changes in stability of cocultures over time. After ~30 generations of evolution, cocultures began occasionally entering phases of very slow growth, but eventually these phases stopped, suggesting that one or both species fixed mutations that would stabilize the coculture. Together, these results show that the *D. vulgaris* and *M. maripaludis* strains we studied have a capacity for rapid evolutionary improvements in syntrophic growth, but the process enabling this change depends on the heterogeneity of the environment, and may cause instability in the community.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Development and Analysis of Multispecies Consortia to Study Microbial Community Stress and Survival

L.D. Miller,[1] A. Venkateswaran,[1] J. Mosher,[1] M. Drake,[1] Z.K. Yang,[1] M. Rodriguez,[1] S.D. Brown,[1] T.J. Phelps,[1] M. Podar,[1] A.V. Palumbo,[1] C.W. Schadt,[1] M. Keller,[1] D.C. Joyner,[2] T.C. Hazen,[2] S. Stolyar,[3] K. Hillisland,[3] and D.A. Stahl[3]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [3]Dept. of Civil and Environmental Engineering, University of Washington, Seattle, Wash.

Cultivation of individual microbial species has been at the core of experimental microbiology for more than a century but offers only a glimpse into the metabolism and ecophysiological potential of most microorganisms. Microbial communities, not individual species, control process rates and drive key biogeochemical cycles, including those that determine the transformation of environmental pollutants of concern to the DOE. Thus controlled studies of model consortia comprised of multiple species that mediate such processes are essential for advancing DOE objectives in bioremediation and other applications.

As part of phase two of the Environmental Stress Pathway Project (ESPP) team, the goals of this project are: (i) to develop a stable syntrophic microbial consortia in continuous flow systems which can be used for physiological and functional genomic studies, (ii) to study how organisms change their activity on a molecular level in response to other community members in the co-culture by analyzing the gene transcription, metabolite flux and growth, and (iii) to study how these changes in activity and community organization affect the resistance and resilience of microbial communities in response to stress, invasion and other perturbations on a molecular level.

The project workflow has two major and interrelated components: microbial cultivation and stress experiments and development and application of analytical methods to characterize the consortia. *Clostridium cellulolyticum* was chosen as the basal organism for initial the three & four member syntrophic assemblies as it can ferment cellobiose, producing acetate, lactate, ethanol and hydrogen. The secondary stage in the syntrophic chain is represented by *Desulfovibrio vulgaris* and by *Geobacter sulfurreducens*, which utilize the *C. cellulolyticum*-produced metabolites. *D. vulgaris* and *G. sulfurreducens* are provided with sulfate and fumarate respectively as electron acceptors. Additional studies with a methanogens (*Methanococcus maripaludis* and other species) are also being pursued. Methods for tracking population dynamics such as quantitative PCR and species specific fluorescently labeled antibodies have been developed and have shown that stable assemblages comprised of these species can be achieved. The chemostat instrumentation operates multiple fermenters that are fed from the same medium source. The multiple fermentation setup can be used for providing biological replicates as well as for conducting stress experiments. Methods for quantitative metabolite analysis via HPLC, GC/MS have been successfully developed and the data obtained from analytical methods are consistent with the expected growth and metabolite uptake and depletion patterns for the consortium.

Our initial metabolic analyses of the trimember consortia indicates that growth of *C. cellulolyticum* is carbon limited by the cellobiose concentration. However, its fermentative product acetate remains in abundance despite serving as a carbon source for *D. vulgaris* and a carbon and electron source for *G. sulfurreducens. D. vulgaris* likely relies on hydrogen for an electron source and the small amount of lactate from *C. celluloyticum* and is thereby growth limited by electron donor potential. Growth of *G. sulfurreducens* is limited by abundance of it's electron acceptor, fumarate. In experiments with the methanogen (*M. maripaludis*) added to the consortia it's populations are inherently less stable and hover at levels nearing the detection limit of the QPCR assay, although methane can often be detected at low levels suggesting biological activity of these organisms. After achieving this stable configuration, we have recently developed a four species microarray for tracking transcriptional activity of these species and comparisons with parallel mass spectrometry-based proteomics. We have begun analyzing data from growth experiment designed to compare stable consortia conditions contrasted with individual species grown alone, and are planning near term experiments to understand resistance and resilience of these communities under various perturbation scenarios on a molecular level.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Characterization of Metal-Reducing Communities and Isolates from Uranium-Contaminated Groundwater and Sediments

B. Ramsay,[1,8] C. Hwang,[1,8] S. Carroll,[2] A. Lapidus,[3] J.C. Detter,[3] C. Han,[3] M. Land,[3] L. Hauser,[3] T.C. Hazen,[4,8] **A. Arkin**[4,8] (APArkin@lbl.gov), A. Belieav,[5] R. Sanford,[6] F. Löeffler,[7] and M.W. Fields[1,8]* (matthew.fields@erc.montana.edu)

[1]Montana State University, Bozeman, Mont.; [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]DOE Joint Genome Institute, Walnut Creek, Calif.; [4]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [5]Pacific Northwest National Laboratory, Richland, Wash.; [6]University of Illinois; [7]Georgia Institute of Technology, Atlanta, Ga.; and [8]Virtual Institute for Microbial Stress and Survival

Project Goals: A goal of VIMSS is to characterize the responses of bacterial populations at multiple levels of resolution in order to understand biochemical capacity at DOE waste sites.

The elucidation of how metal-reducing bacteria interact in a given community and how the community responds to stress and perturbations can help infer the interplay between stress pathways and gene networks that help optimize ecosystem function and stability. A goal of VIMSS is to characterize the responses of bacterial populations at multiple levels of resolution in order to understand biochemical capacity at DOE waste sites. Within this context, bacteria with desired functions (e.g., heavy metal reduction) have been isolated from the Oak Ridge Field Research Center (ORFRC). The ORFRC is located within the Y-12 Security Complex near Oak Ridge, TN in the Bear Creek Valley, and the site includes 243-acres of a previously disturbed contaminated area. The subsurface at the FRC contains one of the highest concentration plumes of mobile uranium located in the United States, and contains various levels of nitrate, heavy metal, and organic contamination (http://www.esd.ornl.gov/orifrc/).

Recently, biostimulation with ethanol was used to detoxify the contaminated groundwater. The experiment successfully reduced nitrate and uranium levels to a safe-to-humans level during the span of the trial. During the biostimulation, d-*Proteobacteria* were detected to predominate the subsurface groundwater, and sequences indicative of the genera *Desulfovibrio*, *Geobacter*, and *Anaeromyxobacter* were observed. Two isolates were achieved, *Anaeromyxobacter* fw109-5 and *Desulfovibrio* FW1012B. The *Anaeromyxobacter* genome has been sequenced at JGI, and the *Desulfovibrio* genome is underway.

Desulfovibrio FW1012B was isolated from well FW101-2B in the bio-stimulation zone during U(VI) reduction at FRC. The isolate can reduce sulfate and utilize pyruvate, fumarate, maleate, lactate, and 1,2-propanediol. Cr(VI) is reduced, and nitrate is reduced without growth. The isolate can utilize triethylphosphate, metaphosphate, and trimetaphosphate as a phosphorus source. The closest cultivated relative is *Desulfovibrio carbinoliphilus* based upon the SSU rDNA gene sequence, and the closest relative with a completely sequenced genome is *Desulfovibrio fructosovorans*. (that was already said above)

Anaeromyxobacter fw109-5 is a mesophilic, iron-reducing bacterium that was isolated from groundwater that had a pH of 6.1 and contained approximately 1.4 mM nitrate and 0.9 μM hexavalent uranium. *Anaeromyxobacter* species are high G+C d-*Proteobacteria* related to the genus *Myxococcus*. Based upon SSU rRNA gene sequences, the closest cultivable relative is *Anaeromyxobacter dehalogenans* 2CP-C with 96.5% sequence identity. The strain fw109-5 grows in the pH range of 4.0 to 9.0, but optimal growth is observed from pH 7.0 to 8.0. To date, known electron donors include acetate, lactate, ethanol, and pyruvate, and electron acceptors include nitrate and iron(III) but not AQDS. Yeast extract and peptone do not support growth, and the organism requires low substrate concentrations for growth (i.e., oligotrophic conditions). Optimal growth occurs under anaerobic conditions, and microaerophilic conditions can be tolerated. The *Anaeromyxobacter* fw109-5 genome is approximatley 5.3 Mb in size with 4,336 candidate protein-coding genes. The slow-growing bacterium is predicted to have two SSU *rrn* genes (16S), and almost 30% of the predicted ORFs are classified as conserved hypothetical proteins. A large percentage of estimated ORFs are predicted to be part of a signal transduction pathway with enrichment in serine/theronine kinase putative proteins. In comparison, fw109-5 had similar numbers of putative two-component and one-component signal transduction proteins as other sulfate- and metal-reducing d-*Proteobacteria*, but fewer compared to *Myxococcus xanthus*. In addition, preliminary data suggest social behavior and sporulation. The genome is predicted to encode a full glycolytic and tricarboxylic acid cycle as well as a pyruvate dehydrogenase complex. Approximately 105 putative proteins are predicted to contain heme-binding sites, with almost half being multi-heme proteins.

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

GTL

Changing Patterns of Selection on γ-Proteobacteria Revealed by the Ratio of Substitutions in Slow:Fast-Evolving Sites

Jesse Shapiro[1] and **Eric Alm**[1,2*] (ejalm@mit.edu)

[1]Massachusetts Institute of Technology, Cambridge, Mass.; and [2]The Virtual Institute for Microbial Stress and Survival, Berkeley, Calif.

Project Goals: This work is consistent with the DOE stated goal "to gain insights about fundamental biological processes"

Different microbial species are thought to occupy distinct ecological niches, subjecting each species to unique selective constraints, which may leave a recognizable signal in their genomes. Thus, it may be possible to extract insight into the genetic basis of ecological differences among lineages by identifying unusual patterns of substitutions in orthologous gene or protein sequences. We use the ratio of substitutions in slow versus fast-evolving sites (S:F) to quantify deviations from the typical pattern of selective constraint observed across bacterial lineages. We propose that elevated S:F in one branch (an excess of slow-site substitutions) can indicate a functionally-relevant change, due to either positive selection or relaxed constraint. In a genome-wide comparative study of γ-proteobacterial proteins, we find that cell-surface proteins involved with motility and secretion functions often have high S:F ratios, while information-processing genes do not. Change in evolutionary constraints in some species is evidenced by increased S:F ratios within functionally-related sets of genes (e.g. energy production in *Pseudomonas fluorescens*), while other species apparently evolve mostly by drift (e.g. uniformly elevated S:F across most genes in *Buchnera spp.*). Overall, S:F reveals several species-specific, protein-level changes with potential functional/ecological importance. As microbial genome projects yield more species-rich gene-trees, the S:F ratio will become an increasingly powerful tool for uncovering functional genetic differences among species.

GTL

Natural Diversity and Experimental Evolution of Environmental Stress Tolerance in Marine Bacteria

A.C. Materna,[1*] S.A. Clarke,[1] C. Cruz,[1] X. Gao,[1] and **E.J. Alm**[1,2,3] (ejalm@mit.edu)

[1]Massachusetts Institute of Technology, Cambridge, Mass.; [2]The Virtual Institute for Microbial Stress and Survival, Berkeley, Calif.; and [3]The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Mass.

Project Goals: This work is consistent with the DOE stated goal "to gain insights about fundamental biological processes".

Genome sequencing has revealed extensive genetic variation within bacterial species and among co-existing bacteria. Using marine *Vibrio* strains as a model system, we investigate to what extent observed sequence diversity corresponds to measurable differences in salinity and temperature tolerance phenotypes, two ecologically important factors for this group of organisms. Using directed evolution, we quantify how malleable these phenotypes are with respect to a small number of mutation events. We have designed two-dimensional gradients in 24 cm square dishes containing solid growth medium to monitor temperature and salinity tolerances over a broad range of both factors. Growth patterns indicate the strain-specific minimum and maximum tolerances and interactions between the factors (salinity and temperature). We compared the specific boundaries of growth for multiple strains of *Vibrio splendidus* and *V. alginolyticus*. While the obtained profiles differ in their shape and limits, some consistent features appear. Tolerance to increasing salinities correlated positively with temperature tolerance. However, higher salinity constrained the limits of temperature tolerance, so that the maximum salinity tolerance occurred at intermediate temperatures. Similarly, growth at higher temperatures led to a tradeoff, limiting the range of salinity tolerance. Interestingly, at high salinities, low temperatures tended to suspend growth, leaving viable cells that could be regenerated when the temperature gradient was removed, while higher temperatures led to killing. In addition to comparing related environmental isolates, this method was further applied to study differences between parental and evolved strains. Serial application of 10^6 cells/cm^2 to the solid medium gradients enabled selection for spontaneous, more tolerant mutants. Five rounds of this method has repeatably produced mutants that tolerate salinity levels that kill the ancestor. In future work, this integrated ecological and experimental approach will be combined with genome re-sequencing to draw connections between genetic diversity and ecologically relevant phenotypes and tradeoffs.

* Presenting author

The *Shewanella* Federation

GTL

Constraint-Based Metabolic Modeling Reveals High Maintenance Energy Requirement for Growth of *Shewanella oneidensis* MR-1 Under Elevated O_2 Tensions

Grigoriy E. Pinchuk[1]* (grigoriy.pinchuk@pnl.gov), Jennifer L. Reed,[2] Oleg V. Geydebrekht,[1] Eric A. Hill,[1] Alexander S. Beliaev,[1] and **Jim K. Fredrickson**[1]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.; and [2]Dept. of Chemical and Biological Engineering, University of Wisconsin, Madison, Wis.

Project Goals: The research proposed in this objective is aimed at improving our understanding of the complex nature of energy conversion networks in *Shewanella* including electron acceptor and electron donor utilization pathways. In that regard, we will emphasize providing a system-level conceptual model linking the ETN and the CCM/PCM pathways in MR-1 that will reflect the ecophysiology of *Shewanella* in redox stratified environments.

Shewanella species belong to a diverse group of bacteria that live at redox interfaces present in the different environments including water column, sediments and subsurface. The properties of their habitats have shaped *Shewanella* metabolic flexibility allowing these microorganisms to effectively respond to the challenge of rapid changes in nutrient and electron acceptor type and concentration. As part of the Genomics:GTL funded *Shewanella* Federation consortium, which conducts integrated genome-based research of *Shewanella* ecophysiology, we have studied the metabolic response of *S. oneidensis* MR-1 to changes in O_2 concentration during steady-state aerobic growth.

A constraint-based metabolic model of *S. oneidensis* was used to estimate ATP maintenance parameters for MR-1 grown under aerobic conditions on lactate in a chemostat with high (20% air saturation) dissolved oxygen tension (DOT). Using lactate uptake rate measurements and dilution rates as constraints, we calculated the maximum amount of ATP that could be hydrolyzed (V_{ATP}^M, mmol ATP hydrolyzed / g of dry weight biomass / hr). A linear relationship between cellular growth rate and maximal ATP hydrolyzation was found:

$$V_{ATP}^M \text{ (mmol/gDW/hr)} = 1.2 + 225 \cdot \mu \qquad \text{(Eq. 1)}$$

Based on these values, the non-growth associated ATP maintenance (NGAM) term was 1.2 (mmol/gDW/hr), which is similar to values for other organisms: *Escherichia coli* (7.6), *Geobacter sulfurreducens* (0.45), and *Bacillus subtilis* (9). However, the growth associated ATP maintenance (GAM) term (225 mmol/gDW) was significantly higher than has been reported for other bacteria: *E. coli* (~46), *G. sulfurreducens* (~47), and *B. subtilis* (~104).

These ATP estimates were consistent for six different dilution rates when the DOT in the reactor was high (note that only four dilution rates were used in fitting the V_{ATP}^M equation). We also calculated the V_{ATP}^M for conditions under O_2 limitation and low DOT, and found that the maximal amount of ATP that was hydrolyzed under these conditions was much lower (about 3-fold) than would be estimated using the equation 1 above. Several hypotheses were generated to explain the high ATP hydrolysis rates observed under elevated O_2 concentrations: (i) oxidative damage of redox proteins enhances their intracellular turnover rate; (ii) *S. oneidensis* MR-1 forms aggregates containing a major extracellular polymeric substance (EPS) component in the presence of mmolar concentration range of Ca^{2+} ions under elevated O_2 levels. Cell aggregation is thought to be a defense mechanism against oxidative stress. DNA, carbohydrates, and proteins are known to be an important part of EPS in these aggregates. We grew bacteria in a medium that contained trace $CaCl_2$, therefore no aggregation was observed although presumably EPS was still produced. We hypothesize that the extracellular polymers produced under highly aerobic conditions where hydrolyzed and then recycled by growing cells thus generating a futile cycle reflected in elevated ATP spending; (iii) oxidative stress caused by high DOT triggered cell lysis possibly as a result of lysogenic conversion since *S. oneidensis* MR-1 chromosome contains DNA encoding for 3 prophages. Experimental assessment of these hypotheses is underway, however first TEM images taken for high O_2 cultures showed significant amounts of viral particles present in *S. oneidensis* growth medium.

This work represents an important step toward understanding of *Shewanella* species ecophysiology using constraint-based analysis as a key tool for investigating energy metabolism.

GTL

Phylogenetic Footprinting in the *Shewanellae*

Lee Ann McCue[1]* (leeann.mccue@pnl.gov), William A. Thompson,[2] Lee A. Newberg,[3,4] and **Charles E. Lawrence**[2]

[1]Pacific Northwest National Laboratory, Richland, Wash.; [2]Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, R.I.; [3]The Wadsworth Center, New York State Dept. of Health, Albany, N.Y.; and [4]Dept. of Computer Science, Rensselaer Polytechnic Institute, Troy, N.Y.

Project Goals: The major goal of this project is the development of the computational bioinformatics tools neces-

sary for study of transcription regulation in prokaryotes by comparative genomics.

Species of the genus *Shewanella* have diverse metabolic capabilities, including their ability to reduce various terminal electron acceptors, leading to considerable interest in their potential for remediation of contaminated environments and use in microbial fuel cells. This interest has spurred the genome sequencing and concomitant comparative genome analysis of over 20 species of *Shewanella*, each with distinct physiology and isolated from various natural habitats (Fredrickson, et al., 2008). As with other comparative genomics studies, the study the transcription regulatory mechanisms of the *Shewanellae* has been facilitated by the availability of these genome data, while simultaneously being confounded by the close phylogenetic relationship of these species. Specifically, while the identification of transcription factor binding sites and regulons can be predicted by analysis of orthologous promoter data (*i.e.*, phylogenetic footprinting), the recent speciation of closely related genomes results in correlation in the sequence data, which confounds the detection of functionally conserved sequence motifs by conventional motif-finding methods.

To address this challenge, we have developed a version of the Gibbs recursive sampler that incorporates phylogeny of the input sequences through the use of an evolutionary model, and calculates an ensemble centroid motif solution (Newberg, et al., 2007); these extensions yield an improved ability to avoid false positive motif prediction in phylogenetically correlated data. We have applied these advanced features of the phylogenetic Gibbs sampler to predict regulatory sites in orthologous data from 17 *Shewanellae*. Furthermore, because many of the predicted sites are likely bound by a common transcription factor, motifs with similar patterns were clustered to infer sets of co-regulated genes (*i.e.*, regulons).

References

1. Fredrickson, J. K., Romine, M. F., Beliaev, A. S., Auchtung, J. M., Driscoll, M. E., Gardner, T. S., et al. (2008). Towards environmental systems biology of *Shewanella*. *Nat Rev Microbiol, 6*(8), 592-603.
2. Newberg, L. A., Thompson, W. A., Conlan, S., Smith, T. M., McCue, L. A., & Lawrence, C. E. (2007). A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, 23(14), 1718-1727.

Constraint-Based Modeling of Metabolism in *Shewanella oneidensis* MR-1

Jennifer L. Reed[1]* (reed@engr.wisc.edu), Sharon J. Wiback,[2] Iman Famili,[2] Jessica De Ingeniis,[3] and Grigoriy Pinchuk[4]

[1]Dept. of Chemical and Biological Engineering, University of Wisconsin, Madison, Wis.; [2]Genomatica, Inc., San Diego, Calif.; [3]Burnham Institute for Medical Research, La Jolla, Calif.; and [4]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

Genome-scale network reconstructions account for components and component interactions in biological networks, and are a way in which to collate and analyze data from a variety of sources. Here we report a metabolic reconstruction for *Shewanella oneidensis* MR-1 based on the current genome annotation and primary literature. The reconstruction includes 760 reactions, 780 genes, and 623 metabolites. A biomass equation specific for *S. oneidensis* has been developed based on published data, and phospholipid and amino acid composition measurements taken under aerobic conditions.

The reconstruction was used to build a constraint-based model that was used in a variety of computational analyses, including: assessment of growth phenotypes, evaluation of metabolite usage (as substrates or by-products), and prediction of knock-out phenotypes to look at metabolic robustness. The model correctly predicted growth on a variety of carbon and nitrogen sources. In addition, quantitative evaluation of alternative electron acceptors led to the identification of four classes of electron acceptors, with differing biomass yields (g D.W. produced per mmol electron acceptor consumed). Gene deletion simulations under different environmental conditions with various carbon sources and electron acceptors identified a large group of genes were never essential (540 out of 780), and a smaller fraction that were always essential (200 out of 78) for growth.

The developed model is a resource that can be used by *Shewanella* researchers and illustrates how reconstructions can serve as a means to evaluate experimental data and generate testable hypotheses to better understand its ecophysiology. This project is a part of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of respiratory-versatile members of this important genus.

Acknowledgements
The co-authors wish to thank all members of the *Shewanella* Federation, especially Margaret Romine, Gretta Serres, and Andrei

Osterman for their work on the genome annotations for MR-1 and other *Shewanella*, which was critical for the development of the metabolic model.

GTL

Progress in Identification of the 'Mobilome' Associated with 21 Sequenced *Shewanella* sp.

Margaret Romine* (Margie.romine@pnl.gov)

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: The *Shewanella* Federation project has four overarching goals: 1) elucidate the environmentally-controlled signal transduction and transcriptional regulatory systems crucial for the ecological success of *Shewanella*, 2) characterize the electron transfer networks and central metabolic pathways involved in energy conversion in *Shewanella*, 3) develop an integrated model of *Shewanella* metabolism, sensing, & regulation, and 4) develop an understanding of *Shewanella* population genomics: understanding evolution, ecophysiology and speciation. In this presentation we will highlight our initial findings on the contribution of laterally acquired DNA to the evolution of this species.

Mobile elements play an important role in the evolution of microbial genomes through activities such as horizontal gene transfer, gene disruption, gene expression modulation, and recombination. There are currently 19 complete and 4 partial *Shewanella* genome sequenced derived from strains that vary considerably in phylogenetic type, environmental origin, and culture conditions (e.g., temperature, salinity, carbon and energy sources) that will support maximal growth rates. These genomes are rich in mobile genetic elements including insertion sequences, transposons, bacteriophage, plasmids, miniature inverted-repeat transposable elements (MITEs), group II introns, integrative conjugative elements (ICE), and mobile genomic islands. This combined richness in mobile elements and broad diversity in strain type that has been sequenced provides an excellent resource for studying the evolutionary events that have enabled members of the Genus to inhabit such a broad variety of niches.

Using comparative sequence analysis the precise termini of many of the mobile elements and consequently to predict integration site specificity and to establish what types of functions have been acquired by different *Shewanella* through lateral transfer. Genome sizes across this group vary from 4.3 Mb (*S. amazonensis*) to 5.9 Mb (*S. woodyii*) and predicted to encode from 3668 to 4941 CDS, respectively. So far, up to 15% of the total predicted protein-encoding genes within a *Shewanella* genome are predicted to be encoded by mobile elements. However, we estimate that the number will be even larger as additional elements are discovered and mapped, including those that are more difficult to delineate due to subsequent evolutionary events that have resulted in attrition or accretion of the originally mobilized element.

Insertion elements are found in all of the genomes and frequently occur within other mobile elements, suggesting that many of the laterally acquired functions do not confer selective advantage to the host and hence their functionality is gradually lost over time. This is not always the case as is exemplified by the large mobile genomic islands that are devoid of IS elements and confer arsenate detoxification and respiratory functionality to *Shewanella* sp. ANA-3 and nitrate assimilatory capability to *S. denitrificans*. The occurrence of genes encoding host-restriction modification enzymes on many of these mobile elements suggest that they also play a significant role in controlling high frequency uptake of DNA from other microbial hosts. Further, the identification of other species that encode genes with high identity to those found in mobile elements of *Shewanella* provide clues as to the types of organisms with which they previously formed close associations with in natural environments and the evolutionary history that resulted in acquisition of new traits or loss of previously encoded ones. For example, the CyaB adenylate cyclase that is encoded in *S. frigidimarina* has between 55 and 64% identity to bi-directional best hits in all of the other sequenced *Shewanella* except *S. denitrificans*, *S. amazonensis*, and *S. woodyii*. However, it possesses slightly better (67-69% identity) similarity to proteins encoded by *Marinomonas*, *Vibro*, and *Allivibrio* and is encoded by mobile elements in both *S. frigidimarina* and *Marinomonas*, while all the remaining *Shewanella cyaB* genes do not appear to be encoded by mobile elements. This finding suggests an independent origin for the *S. frigidimarina cyaB* gene relative to those found in the other sequenced *Shewanella*. The same is true of the NADH-ubiquinone oxidoreductase in *S. oneidensis*. In this case there is currently no clear evidence that it is encoded by a mobile element. However, the absence of genes encoding this function in all of the sequenced *Shewanella* and the observation that they encode proteins with greater than 90% identity to those encoded by *Aeromonas hydrophila* suggests that they share an evolutionary origin and there is a high likelihood that members of these Genera interact with each other or a common partner in natural environments.

We anticipate that as additional genome sequence becomes available from projects that explore broader phylogenetic diversity (e.g. JGI's GEBA project and various metagenome projects) than is currently available, it will be possible to more accurately trace the evolutionary history that resulted in diversification of the *Shewanella* group.

Acknowledgements:
This work was facilitated by use of the Ortholog editor tool in the *Shewanella* Knowledgebase (funded by DOE under the *Shewanella* federation project led by Ed Uberbacher) and the use of neighborhood analysis resources in the JGI Integrated Microbial Genomes (IMG) system and the Fellowship for Interpretation of Genomes SEED system. I would also like to acknowledge LeeAnn McCue for identification of the draft ortholog pairs that were uploaded in the Ortholog Editor, Tatiana Karpinets, Guru Kora, and Denise Schmoyer for development of the Ortholog Editor, and Gretta Serres and Kostas Konstantinidis for identification of strain-specific genes in the first 10 sequenced *Shewanella* genomes.

GTL

Adenylate Cyclases and Anaerobic Respiration in *Shewanella oneidensis* MR-1

M. Charania[1]* (charania@uwm.edu), E. Biddle,[1] Y. Zhang,[2] G. Pinchuk,[2] A. Beliaev,[2] and **D. Saffarini**[1] (daads@uwm.edu)

[1]Dept. of Biological Sciences, University of Wisconsin, Milwaukee, Wis.; and [2]Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: One of our major goals is to identify the networks that are involved in the regulation of anaerobic respiration in *Shewanella oneidensis* MR-1. We are aso interested in identifying the enzymes that are involved in anaerobic respiration of sulfur compounds, and the mechanisms that regulate their expression.

Anaerobic respiration in the metal reducing bacterium *Shewanella oneidensis* MR-1 is regulated by cAMP and its receptor protein CRP. Three *S. oneidensis* proteins, CyaA, CyaB, and CyaC, were shown to have adenylate cyclase activity. Chromosomal deletions of either *cyaA* or *cyaB* did not affect anaerobic respiration with fumarate, DMSO, or Fe(III), whereas deletion of *cyaC* caused deficiencies in respiration with DMSO and Fe(III), but not with fumarate. A double mutant that lacks *cyaA* and *cyaC* was unable to grow anaerobically with the same electron acceptors and exhibited a phenotype similar to that of a *crp* mutant. Microarray analysis of gene expression in the *crp* and *cyaC* mutants indicated that both are involved in the regulation of numerous genes and include the DMSO, fumarate, and Fe(III) reductase genes. Additionally, several genes were differentially expressed in the *cyaC* mutant, but not in the *crp* mutant. Our results indicate that CyaC plays a major role in regulating anaerobic respiration, and may contribute to additional signaling pathways independent of CRP.

In addition to terminal reductase genes, CRP appears to regulate the expression of genes that encode two-component system transduction proteins. These include SO4155 and SO4157. Deletion of these genes in *S. oneidensis* MR-1 led to deficiencies in sulfur, thiosulfate, and tetrathionate reduction. Our results indicate that this deficiency may be due to lack of *phsABC* expression that was predicted to encode the thiosulfate reductase. Mutagenesis of this operon led to significant loss of thiosulfate, tetrathionate, and polysulfide reduction. Thus, it appears that PhsABC is the major enzyme used for the reduction of several sulfur compounds in *S. oneidensis*. This is in contrast to other bacteria, such as *Salmonella typhimurium*, that use different enzymes for the reduction of sulfur compounds.

GTL

Experimental and Computational Analysis of Growth-Phase Dependent Transcriptional Programs in *Shewanella oneidensis*

Qasim K. Beg[1]* (qasimbeg@bu.edu), Niels Klitgord,[2] Mattia Zampieri,[2,3] Timothy S. Gardner,[1,4] and **Daniel Segrè**[1,2,5] (dsegre@bu.edu)

Depts. of [1]Biomedical Engineering, [2]Bioinformatics Program, and [5]Biology, Boston University, Boston, Mass.; [3]International School for Advanced Studies, Trieste, Italy; and [4]Amyris, Emeryville, Calif.

Project Goals: This project is part of the *Shewanella* Federation research activity. As such, it contributes to the overall goal of applying genomic tools to better understanding of the eco-physiology of the respiratory-versatile members of this important genus. Specifically, in our work we expand and improve our knowledge of the physiology of *Shewanella oneidensis* MR-1 by combining measurements of gene expression in time-course experiments under different growth conditions, with computational analysis and flux balance modeling.

In order to study the complex regulatory network of *Shewanella oneidensis* MR-1 and to understand its unique respiratory and metal-reducing abilities we integrate bioreactor experiments under defined environmental/growth conditions with computational analyses of gene expression clusters and metabolic fluxes.

Understanding the complex nature of bacterial responses to changing environmental conditions requires gaining insights into the global regulatory processes at various phases of bacterial growth. The growth-phase related gene expression changes in *E. coli* have been studied in the past; however, not much is known about physiological and transcriptional changes in *S. oneidensis* when cells pass through exponential, stationary, and transition phases. To examine if transcriptional signatures can help understand these changes in cellular organization, we grew *S. oneidensis* MR-1 in both complex (LB) and minimal medium (M4-Lactate). Time-course samples from various phases of *S. oneidensis* growth were subjected to microarray analysis (using *Shewanella* arrays from Affymetrix) for temporal gene expression study along a time-course.

We processed the probe intensities of the microarrays in each data set using standard quality control and normalization methods, and constructed a temporal expression profile for each probe set. To determine the functional modules in each data set, we constructed clusters of probesets by applying the k-means algorithm using the Pearson correlation metric as a distance between expression profiles. We then looked for statistical enrichment of Gene Ontology functions for each cluster. This statistical analysis revealed several clusters showing functional enrichment of specific gene classes during various phases of growth of the bacterium.

We observed very specific gene expression patterns in the generated clusters during various phases of growth (from early- to mid- to late-log, and stationary phase) in both growth media. These patterns include the up-regulation of prophage and flagella/motility related genes in the complex medium (during exponential phase) and minimal medium (upon entry into stationary phase), respectively. We also observed significant up- and down-regulation of energy metabolism, amino acids biosynthesis, fatty acid and phospholipids metabolism genes, as well as stress response and starvation genes. Finally, we show how additional insight into the regulatory program of *S. oneidensis* can be gained by studying the transcriptional organization of genes along the chromosome.

GTL

Metabolic Optimality and Trade-Offs Under Combinatorial Genetic and Nutrient Modifications

David K. Byrne[1]* (Dbyrne@Bu.Edu), Timothy S. Gardner,[2,4] and **Daniel Segrè**[1,2,3]

[1]Graduate Program in Bioinformatics, [2]Dept. of Biomedical Engineering, and [3]Dept. of Biology, Boston University, Boston, Mass.; and [4]Amyris, Emeryville, Calif.

http://prelude.bu.edu

Project Goals: To optimize the metabolic capabilities of *Shewanella oneidensis* and *Escherichia coli*.

Metabolic engineering in microbial hosts for the production of renewable chemicals and energy sources has received considerable attention in recent years. Lactic acid, biohydrogen, and biodiesel are representative examples of renewable chemical commodities that can be produced using biodegradable and sustainable compounds. Yet the production output and efficiency of these commodities still fall significantly short of theoretical limits and are, in some cases, insufficient for practical implementation. Since microbial metabolism is the primary cellular mechanism by which energy is generated and distributed, it is of interest to understand how metabolic pathways may be modulated to increase the proportion of energy diverted for industrial use.

A computational framework that uses models of microbial metabolism has been developed to design genetic and nutrient programs that optimize industrial output. Multiple engineering objectives are considered: maximization of productivity, yield, and/or purity, minimization of economic cost, and combinations thereof. Associated trade-offs and pareto-optimal sets are determined for different formulations. To search for optimal solutions using linear objectives, the strong duality theorem from linear programming is used to formulate a bilevel optimization procedure. For nonlinear objectives, combinatorial and genetic algorithm approaches are implemented.

Optimal predictions indicate that significant improvements in *Escherichia coli* lactic acid synthesis and *Shewanella oneidensis* reducing power and biohydrogen generation are attainable. Thus nutrient and genetic reengineering may be used to synergistically improve bioenergetic output for sustainable chemical commodity production.

GTL

Comparative Analyses Across an Evolutionary Gradient Within the *Shewanella* Genus

Margrethe H. Serres[1]* (mserres@mbl.edu), Kostas Konstantinidis,[2] Margaret F. Romine,[3] Jorge L.M. Rodrigues,[4] Anna Y. Obraztsova,[5] Mary S. Lipton,[3] LeeAnn McCue,[3] James K. Fredrickson,[3] and James M. Tiedje[6]

[1]Marine Biological Laboratory, Woods Hole, Mass.; [2]Georgia Institute of Technology, Atlanta, Ga.; [3]Pacific Northwest National Laboratory, Richland, Wash.; [4]University of Texas, Arlington, Tex.; [5]University of Southern California, Los Angeles, Calif.; and [6]Center for Microbial Ecology, Michigan State University, East Lansing, Mich.

Project Goals: This project is a component of the *Shewanella* Federation and, as such, contributes to the overall goal of applying genomic tools to better understand the ecophysiology and speciation of respiratory versatile members of this important genus.

Shewanella is an environmentally important bacterial genus with a versatile electron accepting capability as well as the ability to degrade a broad range of carbon compounds. Members of the genus have been isolated from diverse geographic locations and habitats such as marine and fresh water columns, sediments, and subsurface environments where they appear to thrive in the redox interfaces. Genome analyses have revealed an expansion of proteins involved in sensing and responding to the environment, including chemotaxis receptors, two-component regulators, outer membrane proteins/receptors, and proteins for synthesis or degradation of cyclic nucleotides (signaling molecules). Some of these genome changes are species specific and may reflect ecological specialization or the speciation process.

Genome sequences have been obtained for over 20 members of the *Shewanella* genus. Strains were selected for sequencing based on their phylogenetic relatedness and include members with similarity at the sub-species level (*Shewanella* sp. MR-4 and MR-7; *S. putrefaciens* W3-18-1 and CN-32) as well as species that are either closely or distantly related. Together the selected genomes represent organisms within a bacterial genus but along an evolutionary gradient of relatedness. This dataset therefore forms a unique framework for studying evolution of microbial genomes.

In this work ten of the *Shewanella* strains have been analyzed for common (core) features representing func-

tions shared among members of this genus. In addition differences in genomic, proteomic and growth patterns have been examined in order to link genotypic differences to phenotypic changes and to speciation. The analyses revealed that genotypic and phenotypic similarities among the organisms could be predicted from their evolutionary relatedness despite evidence of extensive horizontal gene transfer. However, the power of the predictions did depend on the degree of ecological specialization of the organisms studied. Making use of the genetic gradient formed by the ten genomes, we could distinguish ecological effects from evolutionary divergence and rank cellular functions in terms of rates of evolution.

We also detected that differences in whole-cell protein expression patterns for organisms grown under identical conditions were larger than the differences detected at the orfeome level. This suggests that similarity in gene regulation and expression should be included as a parameter when species are described. Overall our results constitute information that may be applied towards understanding bacterial species and genera at a system level. We believe these findings can be transferred to many other ecologically versatile bacteria that are prevalent in important habitats on Earth.

GTL

Evolution of Signal Transduction in a Bacterial Genus

Harold Shanafield* (hshanafi@utk.edu), Luke E. Ulrich, and Igor B. Zhulin

National Institute for Computational Sciences, University of Tennessee, Oak Ridge National Laboratory, Oak Ridge, Tenn.

The recent completion of sequencing of multiple genomes in the *Shewanella* genus provides a unique opportunity to study evolution at a much finer scale than previously possible. Using a bi-directional best BLAST hit approach at the protein domain rather than traditional whole-protein sequence level we analyzed the evolutionary relationships of proteins predicted to be involved in signal transduction. Based on these relationships, we have determined a core set of 99 proteins across the first 11 sequenced *Shewanella* genomes that were highly conserved in both domain architecture and protein sequence. The core included one of the several chemotaxis systems found in *Shewanella* and several two-component regulatory systems. A large group of orthologous signal transduction proteins across multiple genomes showed some primary sequence drift and were classified as "significant similarity", and finally there were several unique signal transduction proteins in each organism. We also quantified a recent disproportionate loss of signal transduction genes in *Shewanella denitrificans* OS217 above and beyond the overall reduction in that organism's genome size, and an enrichment of signal transduction genes in *Shewanella amazonensis* SB2B. Possible relationships of the observed changes with the metabolism and environment are discussed.

GTL

Investigating Environmental Specialization in a Population of *Shewanella baltica* Strains

Jie Deng,[1] Jennifer M. Auchtung[1]* (blackjen@msu.edu), Konstantinos Konstantinidis,[2] and **James M. Tiedje**[1] (tiedjej@msu.edu)

[1]Center for Microbial Ecology, Michigan State University, East Lansing, Mich.; and [2]School of Civil and Environmental Engineering and School of Biology, Georgia Institute of Technology, Atlanta, Ga.

Project Goals: This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics to better understand the ecophysiology and speciation of the environmentally-versatile members of the important genus.

Studying how bacterial strains diverge and how divergence correlates with specialization to new environmental niches is important for understanding the dynamics of environmental communities, which often contain populations of closely related strains, and for understanding the potential fate of genetically engineered microbes in the environment. We are studying these questions using a collection of *Shewanella* strains isolated from different regions within the water column of the Baltic Sea. *Shewanella* species have been intensively studied due to their extensive respiratory versatility and their potential role in bioremediation and microbial power generation. Our analysis centers on a collection of *S. baltica* strains previously shown to occupy characteristic positions within the water column. Physical and chemical analysis of the water column performed at the time of strain isolation revealed the presence of a stable redox gradient and provided additional insights into what environmental variables might be important for specialization. Therefore, these strains were ideal candidates for studying strain divergence and environmental specialization.

We used multi-locus sequence typing to analyze phylogeny of 36 strains and examine how this correlates with position within the water column. We identified 8 distinct clades. One clade showed good correlation with depth of isolation; it contained 86% of strains isolated from the anoxic region of the water column. Strains isolated from other regions of the water column were distributed throughout the other clades, suggesting that if specialization to these other depths had occurred, it had likely not occurred at the genome level. We used comparative genomic hybridization (CGH) to provide additional detail of the genetic diversity among these 36 *S. baltica* strains. Genome sequences from four *S. baltica* strains, isolated from three different depths within the water column and previously sequenced through the DOE-GTL program, were used to design the microarrays. Preliminary analysis of this data revealed that when overall similarities

of the genomes were analyzed using a presence/absence matrix, similar phylogenies were produced to those observed through MLST. However, we also identified a subset of genes that were shared primarily by strains isolated from the anoxic region of the water column. Although the majority of these genes encoded proteins with hypothetical or predicted mobile element functions, one cluster encoded genes predicted to be involved in extracellular polysaccharide biosynthesis. Transcriptional profiling revealed that these putative polysaccharide biosynthetic genes were expressed by one of the strains isolated from the anoxic region, providing additional evidence of the potential importance of these genes. Further characterization of these genes, as well as studies comparing differences in gene expression and physiology among these isolates are ongoing.

GTL

Characterization of *C*-type Cytochromes and Their Role in Anaerobic Respiration in *Shewanella oneidensis* and *S. putrefaciens*

Haichun Gao,[1] Jingrong Chen,[1] Yili Liang,[1]* Soumitra Barua,[1,2] Dongru Qiu,[1] Zhili He[1] (zhili.he@ou.edu), Alexander S. Beliav,[3] Margaret F. Romine,[3] Samantha Reed,[3] Dave Culley,[3] David Kennedy,[3] Yunfeng Yang,[2] James M. Tiedje,[4] James K. Fredrickson,[3] Kenneth H. Nealson,[5] and **Jizhong Zhou**[1,2] (jzhou@ou.edu)

[1]Institute for Environmental Genomics and Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.; [2]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Pacific Northwest National Laboratory, Richland, Wash.; [4]Center for Microbial Ecology, Michigan State University, East Lansing, Mich.; and [5]Dept. of Earth Sciences, University of Southern California, Los Angeles, Calif.

Project Goals: This project is focused applying genome-based experimental and modeling approaches to understanding the ecophysiology of *Shewanella* via four overarching and integrated objectives: 1) Elucidate the environmentally-controlled signal transduction and transcriptional regulatory systems crucial for the ecological success of *Shewanella*; 2) Characterize the electron transfer networks and central metabolic pathways involved in energy conversion in *Shewanella*; 3) Develop an integrated model of *Shewanella* metabolism, sensing, & regulation; and 4) Develop an understanding of *Shewanella* population genomics: understanding evolution, ecophysiology and speciation.

Shewanella strains are renowned for their ability to utilize a wide range of electron acceptors for respiration, which is due to a large number of *c*-type cytochromes in their genome. The arsenal of *c*-type cytochromes is also highly diversified across the 21 sequenced *Shewanella* genomes and only twelve out of the 41 c-type cytochrome of *S. oneidensis* MR-1 are present in all other sequenced strains. Only a few *c*-type cytochromes have been characterized so far. To discern the functions of unidentified *c*-type cytochrome genes in bacterial energy metabolisms, we generated 37 single mutants with in-frame deletion of each individual cytochrome gene in MR-1. Reduction of a variety of electron acceptors was measured and the relative fitness was calculated for these mutants based on competition assays. It was revealed that SO0610, SO1777, SO2361, SO2363, and SO4360 were important under aerobic growth conditions, and that most *c*-type cytochomes play a more important role in anaerobiosis. The *petC* gene appeared to be important to both aerobiosis and ananerobiosis. Our results regarding functions of CymA and MtrC are consistent with previous findings. We also assayed the biofilm formation of these mutants and it was shown that SO4666 might be important for pellicle formation.

S. putrefaciens W3-18-1 lacks orthologues for the secondary metal reductase and accessory proteins (MtrFED) of *S. oneidensis* MR-1. Sputw3181_2446 encodes a decaheme *c*-cytochrome, orthologous to outer membrane primary metal reductase OmcB of MR-1 (60% similarity) while another reductase similar to OmcA in MR-1 was also found in W3-18-1. Sputw3181_2445 encodes an 11-heme *c*-type cytochrome OmcE, which only shares 40% similarity with the decaheme cytochrome OmcA. Single and double in-frame deletion mutants of *omcB* and *omcE* were generated for functional characterization of *omcE* and metal reduction in W3-18-1. Reduction of solid-phase Fe(III) and soluble Fe(III) in *S. putrefaciens* W3-18-1 was mainly dependent on OmcB under anaerobic conditions (with 50 mM lactate as electron donors and Fe_2O_3, α-FeO(OH), β-FeO(OH) and ferric citrate as electron acceptors. W3-18-1 catalyzed a faster reduction of α-FeO(OH) as compared to MR-1, suggesting that other genes might be involved in iron(III) reduction in W3-18-1. As previously observed in MR-1, the deletion of both OMCs (OmcE & OmcB) led to a severe deficiency in reduction of solid-phase Fe(III) in W3-18-1 and the defectiveness was more remarkable in the reduction of soluble iron. The *omcB* and *omcE* genes of W3-18-1 had been expressed with the pBAD vector in *E. coli*. Heme staining assays also demonstrated that the disappearance of specific protein bands in the SDS-PAGE gels were consistent with *omcB* and *omcE* deletion in three mutant samples. These results suggest that *omcE* and *omcB* are actually expressed as cytochrome proteins and could play a central role in metal reduction in *S. putrefaciens* W3-18-1.

GTL

The NapC- and CymA-Dependent Nitrate Reduction in *Shewanella oneidensis* MR-1 and *S. putrefaciens* W3-18-1

Dongru Qiu[1]* (qiu@ou.edu), Haichun Gao,[1] Yili Liang,[1] Jingrong Chen,[1] Zhili He,[1] Margaret F. Romine,[2] Kenneth H. Nealson,[4] James M. Tiedje,[3] James K. Fredrickson,[2] and **Jizhong Zhou**[1] (jzhou@ou.edu)

[1]Institute for Environmental Genomics and Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.; [2]Pacific Northwest National Laboratory, Richland, Wash.; [3]Center for Microbial Ecology, Michigan State University, East Lansing, Mich.; and [4]Dept. of Earth Sciences, University of Southern California, Los Angeles, Calif.

Project Goals: This project is focused applying genome-based experimental and modeling approaches to understanding the ecophysiology of *Shewanella* via four overarching and integrated objectives: 1) Elucidate the environmentally-controlled signal transduction and transcriptional regulatory systems crucial for the ecological success of *Shewanella*; 2) Characterize the electron transfer networks and central metabolic pathways involved in energy conversion in *Shewanella*; 3) Develop an integrated model of *Shewanella* metabolism, sensing, & regulation; 4) Develop an understanding of *Shewanella* population genomics: understanding evolution, ecophysiology and speciation.

Nitrate respiration systems are highly diverse among *Shewanella* species. Bioinformatics analyses revealed three types of nitrate reduction systems in *Shewanella* genomes. The well studied *S. oneidensis* MR-1 harbors only the CymA-dependent nitrate reductase and the NapGH ubiquinol oxidase. Most *Shewanella* species including *S. putrefaciens* W3-18-1 have both CymA- and NapC-dependent nitrate reductases, as well as the NapGH ubiquinol oxidase. The *S. baltica* strains have both the CymA- and NapC-dependent nitrate reductases but lack the NapGH ubiquinol oxidase. *S. oneidensis* MR-1 appear to be atypical because it lacks both *napC* and *nrfBCD*, whose gene products act to transfer electrons from quinol pool to terminal reductases NapA and NrfA. In *E. coli*, NapC and NrfBCD are essential for catalyzing reductions of nitrate to nitrite and the subsequent reduction of nitrite to ammonium, respectively. Our previous results revealed that CymA is likely to be functional replacement of both NapC and NrfBCD in the nitrate and nitrite reduction in *S. oneidensis* MR-1.

The nitrate reduction in W3-18-1 and a closely related strain CN32 were examined, and the transient accumulation of nitrite was also observed in both strains fed with 2mM of nitrate (50 mM of lactate as electron donor), suggesting that the two-step manner of nitrate reduction may be common among *Shewanella* species. A series of mutants in W3-18-1 were generated and tested for the bacterial growth on nitrate. Deletion of entire operon of *nap1* (*napDAGHB*) or *nap2* (*napDABC*) did not significantly affect the cell growth, but the double mutant with deletion of both *nap* operons, *nap1* and *nap2*, could not grow on nitrate. It is suggested that the two *nap* operons are functionally redundant. In addition, the in-frame deletion mutants of *cymA* and *napC* of W3-18-1 did not show severe growth defectiveness on nitrate, though deletion of *cymA* rendered the loss of reduction of nitrate and nitrite and bacterial growth in MR-1. Furthermore, the *cymA* deletion mutant showed little growth on nitrite in contrast to the *napC* deletion mutant, indicating that CymA was involved in nitrite reduction in both W3-18-1 and MR-1. The *cymA* gene from W3-18-1 could complement the MR-1 *cymA* in-frame deletion mutant in reduction of ferric ions, nitrate, and nitrite when expressed *in trans*. The *napC* gene from W3-18-1 could also complement the defectiveness of MR-1 *cymA* deletion mutant in ferric iron reduction but it failed to complement nitrite reduction. These results support the hypothesis that the complicated periplasmic nitrate reduction systems, NapC-dependent and CymA-dependent, represent a scavenge mechanism for nitrate and nitrite in most *Shewanella* species, which could utilize nitrate and nitrite as both electron acceptors for energy generation and nitrogen nutrient source for cellular metabolisms.

GTL

The *Shewanella* Knowledgebase

Tatiana Karpinets[1,2] (karpinetstv@ornl.gov),[3] Margaret Romine,[1] Denise Schmoyer,[1,2] Guruprasad Kora,[1] Mike Leuze,[1] Mustafa Syed,[4] Erich Baker,[1] Byung Park,[3] Andrey Gorin,[1,5] Nagiza Samatova,[1] and **Ed Uberbacher**[1]* (ube@ornl.gov)

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]University of Tennessee, Knoxville, Tenn.; [3]Pacific Northwest National Laboratory, Richland, Wash.; [4]Baylor College of Medicine, Waco, Tex.; and [5]North Carolina State University, Raleigh, N.C.

http://shewanella-knowledgebase.org

Project Goals: Provide a framework for data and knowledge integration and sharing for the *Shewanella* Federation.

The *Shewanella* Knowledgebase is a framework for integrating diverse experimental information on *Shewanella* species with their genomic characteristics. Recent developments in the *Shewanella* Knowledgebase have concentrated in the following areas:

1. Developing Ortholog and Genome Editors for 21 sequenced *Shewanella*

2. Further enhancement of the *Shewanella* Knowledgebase web portal, which is a data and knowledge integration environment that allows investigators to query across the *Shewanella* Federation (SF)

experimental datasets, link to *Shewanella* and other community resources, and visualize the data in a cell systems context.

3. Developing a database of SF presentations and reporting materials

The Ortholog and Genome Editor is a set of comparative visualization and analytical tools for curation, storage and analysis of genome annotation for all sequenced *Shewanella* (Fig A). This tool is unique in that no other databases or web servers provide the environment to curate orthologous groupings of proteins present across a set of related organisms. The groupings are especially important for comparative analysis of the species and for improving genome annotations of newly sequenced organisms, which are closely related to a better annotated and studied model organism. In case of Shewaella spp. the model organism is *Shewanella oneidensis MR-1*, which has significant number of orthologous genes with other species of the genus. By developing the ortholog editor we have achieved the following goals: (i) decreased errors generated by automatic prediction of ortholog, (ii) identified genes, which were missed in the initial annotation, and pseudogenes, (iii) provided environment to store, edit and download the most recent annotations of the sequenced *Shewanella* genomes in different formats, (iv) improved product annotation and made it consistent across all *Shewanella* genomes, (v) make it possible to compare omics data at the gene and pathway level across multiple genomes. In addition to the ortholog table the editor database includes a table of evidence supporting the automatic identification of orthologs. The table includes a set of annotations of the genomes of sequenced *Shewanella* spp using different databases and tools including PRotein K(c) lusters, COG, CDD, Pfam, and SMART. Automatic checks of consistency are implemented for locus tags combined into one orthologous group considering both the length of proteins and their domain structures. The main page of the ortholog editor provides access to the ortholog table, capability for results checking, and the generated alignments (Fig. B). A set of options for table editing, sorting, viewing, searching and downloading are also available. The ortholog editor is linked to individual spp, editors that provide a web interface to store, edit and download in different formats the improved annotations of the sequenced *Shewanella* genomes.

The Shewanella Knowledgebase web portal was advanced by collecting experimental data produced by the SF and other researchers with subsequent integration of the data in the knowledgebase environment. A visualization and analytical tool, ShewRegDB, was developed for integration of the regulatory information with metabolic pathway predictions from ShewCyc and with collected experimental information. The Knowledgebase was supplemented by several experimental projects including 3 projects on protein structure and 25 projects involving Affymetrix microarray and two-color microarrays profiling. The regulatory database was enhanced by including 359 binding sites extracted from literature, by independent transcription terminator predictions from TransTermHP, by computational predictions of operons from the DOOR database, MicrobesOnline, ODB, Shewybase and by manually curated operon annotations. An interface to the regulatory database was developed to search three main data objects: DNA regulator binding sites, RNA regulators and operons. The interface links the regulatory information to information stored in the knowledgebase and provides table overviews and downloads of the search results in EXCEL format. Visualization of the regulatory information is now integrated with the experimental data collected in the knowledgebase and pathway predictions in the updated version of ShewCyc. This information can be displayed through additional tracks in the GBrowser. A Metabolic Pathway Viewer based on Google map technology is under development to track metabolites, overlay experimental data with the metabolic pathways and provide bird eye view of each individual pathway.

Database of SF members' progress reports was created providing the following main options: (i) a quick overview of the available reports, presentations, abstracts and posters generated by federation; (ii) downloading all types of reporting materials; (iii) a user friendly interface for uploading a report/presentations/abstracts/poster.

This project is a component of the *Shewanella* Federation and contributes to revealing molecular functions underlying the diverse ecophysiology of this important species.

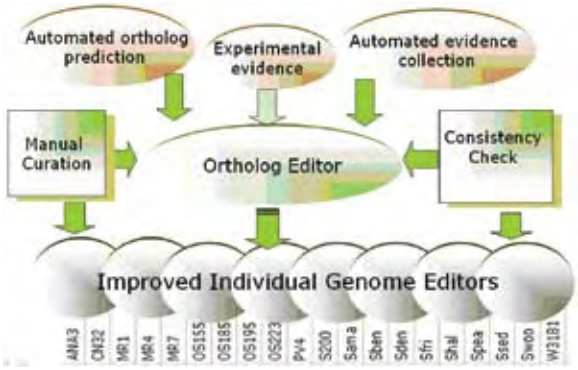


Figure 1: (A)

Figure 2: (B)

Geobacter Systems Understanding

GTL

Experimental Genome Annotation of *Geobacter sulfurreducens*

Karsten Zengler,[1] Byung-Kwan Cho[1]* (bcho@ucsd.edu) Yu Qiu,[1] Christian Barrett,[1] Young-Seoub Park,[1] Mary Lipton,[2] **Derek Lovley**[3] (dlovley@microbio.umass.edu), and Bernhard O. Palsson[1]

[1]Dept. of Bioengineering, University of California San Diego, La Jolla, Calif.; [2]Biosciences Division, Pacific Northwest National Laboratory, Richland, Wash.; and [3]Dept. of Microbiology, University of Massachusetts, Amherst, Mass.

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

The ability to obtain sequence information and assemble whole genomes, or large parts of it, has been paramount to our understanding of organisms, ranging from bacteria to eukaryotes including humans. Recent advance in sequence technologies have improved the quality, quantity, the time frame in which data can be generated and the cost in ways that have been unimaginable just a few years ago. Generating large amounts of high quality sequence data therefore is not a limiting step in biological sciences any longer. However, when it comes to translating sequences into genes, determine genome structure, and assigning function we rely exclusively on computational methods. Precise annotation of every gene within an organism solely by bioinformatics tools has still not been possible.

Realizing this shortfall there have recently been approaches to experimentally annotate genomes by studying the transcriptome of organisms using high density microarrays and sequencing approaches, respectively. Transcriptome analysis by itself, however, can only answer indirectly what the functional elements of the genome (transfrags) are and how these elements are organized and regulated—information which is crucial for comprehensive genome annotation. Here we describe the use of high-density tiling microarrays to determine transcripts abundances, novel transcripts, transcription units, and UTR boundaries under different growth conditions. In addition we determined genome-wide binding locations of DNA-binding proteins including RNA polymerase and sigma factors by ChIP-chip. High-density tiling microarrays of overlapping probes can effectively address the transfrags determination since they are constructed without any *a priori* knowledge of the possible transfrag content of a genomic sequence. However, for an experiment-based annotation it is still necessary to pinpoint the accurate structures (i.e., the boundaries of UTRs and transcription start site) of transcription units (TU). In order to map the transcription start sites of the bacterial genomes, we developed a modified mRNA-seq method using RNA ligation followed by Solexa sequencing. Whether novel transcripts were encoding for proteins or maybe non-coding RNAs was addressed by comprehensive proteomic analysis (>85% coverage of formerly predicted coding proteins). Integration of all these genome wide data sets allowed for a comprehensive experimental genome annotation including several new transcripts and non-coding RNAs.

GTL

Novel Approaches for Genome-Scale Spatial Analysis of Gene Transcription in Biofilms: Elucidation of Differences in Metabolism Throughout *Geobacter sulfurreducens* Biofilms Producing High Current Densities

Ashley E. Franks* (aefranks@microbio.umass.edu), Kelly P. Nevin, Kengo Inoue, Richard H. Glaven, and **Derek R. Lovley**

Dept. of Microbiology, University of Massachusetts, Amherst, Mass.

http://www.geobacter.org

Project Goals: The goal of this project is to provide computational tools to predicatively model the behavior of two microbial communities of direct relevance to Department of Energy interests: 1) the microbial community responsible for in situ bioremediation of uranium in contaminated subsurface environments; and 2) the microbial community capable of harvesting electricity from waste organic matter and renewable biomass. Based on our current rate of progress, it is expected that within five years it will be possible to predict the in situ growth and activity of the relevant microorganisms in the environments of interest solely from relevant geochemical data and to describe in detail the in situ metabolic state of the microorganisms from environmental gene expression data. Furthermore, these computational tools will be able to predict the response of the microbial community to environmental manipulations

or manipulation of the genome of the relevant organisms, allowing rational optimization of in situ uranium bioremediation or electricity harvesting via environmental or genetic engineering.

Microbial fuel cells (MFCs) show promise as a strategy for converting wastes and biomass directly to electrical current, providing a new form of bioenergy. Studies in a number of laboratories have demonstrated that, in the most highly effective MFCs *Geobacter sulfurreducens*, or closely related organisms, outcompete the other diverse constituents of complex microbial communities to preferentially colonize the current-harvesting anodes. Therefore, further information on the growth and metabolism of *G. sulfurreducens* on anodes is expected to lead to the design of superior MFCs. Our previous studies have demonstrated that *G. sulfurreducens* produces thick (>50 μm), electrically conductive biofilms on MFC anodes in systems designed for high current densities. It was hypothesized that environmental conditions might vary significantly throughout the biofilm and, in some instances, might result in suboptimal physiological states that could negatively impact on current production.

Therefore, a MFC that permits real time imaging of the anode biofilm with confocal scanning laser microscopy was developed. *G. sulfurreducens* produced current in this new system at rates comparable to that previously reported with other MFC designs. Cells engineered to produce the fluorescent protein mcherry to facilitate real-time imaging produced current comparable to wild-type cells. Metabolic staining of a current-producing biofilm demonstrated metabolic activity across the entire thickness of the electrically conductive biofilm. Introducing C-SNARF-4, a pH-sensitive fluoroprobe, into the anode chamber revealed strong pH gradients within the anode biofilms. The pH decreased with increased proximity to the anode surface and from the exterior to the interior of biofilm pillars. Near the anode surface pH levels were as low as 6.1 compared to ca. 7 in the external medium. Various controls demonstrated that the proton accumulation was associated with current production. These results demonstrated that it is feasible to non-destructively monitor the activity of anode biofilms in real time and suggest that the accumulation of protons that are released from organic matter oxidation within anode biofilms can limit current production.

In order to investigate levels of gene transcription throughout the biofilm, reporter plasmids were created by cloning short half-life fluorescent proteins into the plasmid pRG5. Promoters of genes of interest were inserted upstream of the fluorescent proteins, controlling their expression. This makes it possible to track gene expression at different layers of the biofilm in real time with confocal microscopy. The short half-life of the fluorescent proteins permits monitoring of repression of gene expression as well as induction. Initial studies are focusing on expression of the gene for the enzyme citrate synthase because our previous studies with chemostat-grown *G. sulfurreducens* have demonstrated that expression levels of this gene are related to rates of metabolism. Other reporters will include those for the gene, *pilA*, which encodes the structural protein for the electrically conductive, microbial nanowires, and for cytochromes that contribute to the electrical conductivity of the biofilm and/or serve as electrical contacts between the cells and the anode.

Another strategy that we developed for investigating gene expression throughout the biofilm involved a novel biofilm sectioning technique coupled with whole-genome microarray analysis of transcript abundance. Biofilms were treated with RNase protect, hardened with resin, and sliced horizontally in 100 nm slices with a diamond knife. Slices were pooled into inner and outer portions, the RNA extracted and transcript abundance analyzed with microarrays. There were146 genes which were differentially expressed (2-fold cutoff; $p<0.05$) between the inner and outer portions of the biofilm. Only 1 gene, encoding a putative ABC transporter, was up-regulated more than 2-fold in the outer portion of the biofilm. Of the genes down regulated in the outer portion of the biofilm, many had to do with cell metabolism and growth, consistent with the concept that cells at a distance from the anode have lower rates of metabolism.

Recent studies on the outer-surface *c*-type cytochrome, OmcZ, illustrate how such gene expression studies can lead to enhanced understanding of the functioning of *G. sulfurreducens* in anode biofilms. Microarray analysis indicated that *omcZ* was much more highly expressed in current-producing biofilms than in biofilms growing on the same electrode material, but using fumarate as the electron acceptor. Gene deletion and complementation studies demonstrated that *omcZ* was essential for high-density current production. Electrochemical analysis demonstrated that the biofilms of the *omcZ*-deficient strain were highly resistive. OmcZ was purified and characterized. Preliminary localization studies with gold-labeled antibodies demonstrated that OmcZ was extracellular and dispersed throughout *G. sulfuureducens* biofilms. These results suggest that OmcZ enables long-range electron transfer through *G. sulfurreducens* anode biofilms by providing a mechanism for electron transfer through the biofilm matrix. If so, then increasing expression of OmcZ may improve biofilm conductivity and hence MFC current densities.

These studies represent the first analysis of the gene expression and physiological status of high current density biofilms. The microtoming studies represent the first genome-scale analysis of gene transcript abundance within individual depths of a biofilm of any type. The results from these studies are providing new insights into the function of MFC anode biofilms that are expected to lead to better strategies for optimizing the power output of MFCs and thus broaden their applications.

GTL

Genome-Wide Mapping of Transcriptional Start Sites of *Geobacter sulfurreducens* using High-Throughput Sequence Methodologies

Katy Juárez[1]* (katy@ibt.unam.mx), Alfredo Mendoza,[1] Leticia Vega-Alvarado,[2] Blanca Taboada,[2] Verónica Jimémez,[3] Leticia Olvera,[1] Maricela Olvera,[1] Ana L. Tirado,[1] Julio Collado-Vides,[3] Enrique Morett,[1] and **Derek R. Lovley**[4] (dlovley@microbio.umass.edu)

[1]Instituto de Biotecnología, [2]Centro de Ciencias Aplicadas y Desarrollo Tecnológico and [3]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico; and [4]Dept. of Microbiology, University of Massachusetts, Amherst, Mass.

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

U(VI) bioremediation strategies require an extensive understanding of metal-reducing microorganisms, the environmental parameters controlling their metabolism and the regulation of their gene expression. Molecular analyses have clearly indicated that *Geobacteraceae* are the predominant bacteria of the microbial community during in situ bioremediation of uranium-contaminated environments and when dissimilatory metal reduction is stimulated in the subsurface. Although significant progress has been made in the knowledge of the electron transfer mechanisms in *Geobacter sulfurreducens*, little is known about the regulatory cues involved in controlling gene expression of the participants in this complex process.

Gene expression starts by the specific and normally tightly regulated transcription initiation process. The experimental determination of the genome-wide transcription start sites (TSS), will contribute, not only to improve the knowledge about the promoter and operon structure in this bacterium, but also to increase the predictive capacity for those cases not experimentally determined. Having precise information about the majority of the TSSs in *G. sulfurreducens* will certainly help to achieve a detailed understanding of the global regulatory circuits that control gene expression in this organism.

In order to obtain a global picture of the active promoters in *G. sulfurreducens* growing in different environmental conditions, we have developed two approaches to map thousands of transcriptional start sites: one is directed mapping using a modification of the 5'RACE protocol, and the other one is a global mapping of TSS using 454 pyrosequencing technology. Using these strategies in *G. sulfurreducens*, we have been able to map hundreds of new TSS and confirm some TSS previously reported, indicating that our methodology is accurate and robust. We also developed a web-based tool to analyze the results of the pyrosequencing reactions and the information obtained is available at http://geobactertss.ccg.unam.mx and it has allowed us to analyze very efficiently the individual sequences over the *G. sulfurreducens* genome.

The results of the first cDNA library show a distribution throughout the genome and in correlation with the gene expression levels. We are evaluating the possible occurrence of intragenic promoters and new coding small ORFs. Interestingly, in this first collection we identified additional TSS for some genes previously studied. This work adds substantial new data to our ongoing large-scale effort to experimentally determine TSS, promoters and regulatory elements in *G. sulfurreducens*. We have found that it is not uncommon that genes are expressed from multiple promoters, which certainly enriches the regulatory options to control gene expression. Our results provide new information that certainly will facilitate the understanding from a global perspective the complex and intricate regulatory network that operates in *G. sulfurreducens*.

GTL

Genome Resequencing Reveals that Current-Harvesting Electrodes Select for Rare Variant of *Geobacter sulfurreducens* Capable of Enhanced Current Production

Anna Klimes[1]* (aklimes@microbio.umass.edu), Kelly P. Nevin,[1] Hana Yi,[1] Jessica E. Butler,[1] Trevor L. Woodard,[1] Christian L. Barrett,[2] Karsten Zengler,[2] Bernhard Palsson,[2] Barbara Methe,[3] Yuan Gao,[4] Bin Xie,[4] Chen Liang,[5] and **Derek R. Lovley**[1]

[1]Dept. of Microbiology, University of Massachusetts, Amherst, Mass.; [2]Dept. of Bioengineering, University of California San Diego, La Jolla, Calif.; [3]J. Craig Venter Institute, Rockville, Md.; [4]Virginia Commonwealth University, Richmond, Va.; and [5]Beijing University of Aeronautics and Astronautics, Beijing, China

Project Goals: The goal of this project is to provide computational tools to predictively model the behavior of two

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

Electricity is not commonly considered as a form of bioenergy, but microbial fuel cells (MFCs) can effectively convert a range of organic wastes and biomass directly to current. At present, the application of MFCs is limited by low power outputs. One likely reason for the low power is that electron transfer to electrodes is probably only a fortuitous process, related to extracellular electron transfer to natural extracellular electron acceptors such as Fe(III) oxides. As far as is known, there has been no previous selective pressure on microorganisms to produce current. Thus, this process is probably far from optimized.

Geobacter sulfurreducens is capable of producing current densities as high, or higher, than any other known pure or mixed culture. In an attempt to develop a strain of *G. sulfurreducens* with improved current production capability we repeatedly transferred cells grown on the current-harvesting anodes of microbial fuel cells, or selected for growth at low anode potentials, with the assumption that this selective pressure would result in the accumulation of beneficial mutations that would enhance the capacity for electron transfer to anodes. Strains were obtained that produced up to 30-fold more power than the starting culture.

Resequencing the genomes of twelve of these electrode-selected strains with Illumina sequencing technology revealed that the sequences of the 16S rRNA operons were identical to the starting culture, indicating the electrode-selected strains were *G. sulfurreducens*. However, each strain had over 18,000 nucleotide mutations compared with the genome sequence obtained from the starting culture. Furthermore, the electrode-selected strains appeared to be missing ca. 5% of the sequence found in the starting culture. Similar results were obtained with Nimblegen re-sequencing chip technology. About a third of the genes in the genome contained at least one nucleotide mutation. The mutations were found throughout the length of the genome, with some genes containing over 100 mutations. There was a heavy bias for the mutations to be in the third position of the codon, which resulted in the majority of the mutations (ca. 75%) causing synonymous changes, meaning no change in the protein sequences. Furthermore, the electrode-selected strains contained ca. 100 kb of DNA sequence that was not detected in sequencing of the starting culture.

This degree of genome change is much greater than can be attributed to random mutation and genome rearrangements during the course of selection on the electrodes. Rather, the results suggest that the starting culture of *G. sulfurreducens* contains at least two strains: the strain that predominates under all previously described culturing conditions (referred to here as strain DL1) and at least one variant that is specifically selected during growth on electrodes. The fact that the variant was not detected in the original culture even after 80-fold coverage of the genome sequence with Illumina sequencing indicates that the electrode-specific variant is in very low abundance under normal culturing conditions. The most heavily mutated genes, common to all sequenced strains of the variant, are predicted to have roles in energy metabolism, cell envelope synthesis, and transport. Several TCA cycle proteins and several clusters of cytochromes were heavily mutated, which may affect electron transport pathways. In addition, there were nonsynonomous mutations in proteins required for DNA synthesis and replication, and for DNA mismatch and strand break repair, suggesting that the electrode-selected variants might have a higher mutation rate than the DL1 strain.

PCR primers designed to amplify four of the genes found in the variants, but not in the DL1 strain, amplified the correct PCR product from the original *G. sulfurreducens* culture. However, a nested PCR protocol was required, suggesting that the variant was in very low abundance in the culture. This low abundance was verified with quantitative PCR and is consistent with results of Illumina sequencing of the starting culture.

In addition to their enhanced capability for current production, the electrode-selected variants had other unique features that distinguished them from the DL1 strain. For example, the electrode-selected variants were highly motile, with long flagella, whereas whereas DL1 is non-motile and does not produce flagella. Compared with DL1, the electrode-selected strains had a greater abundance of pilin-like filaments that look like the electrically conductive pili, known as microbial nanowires. Furthermore, analysis of one of the electrode-selected strains demonstrated that its biofilms are much more conductive than DL1 biofilms. Electrochemical analysis of the anode biofilms of this strain demonstrated that it had a lower mid-point potential than DL1 for electron transfer to electrodes, consistent with its ability to produce much higher current on electrodes poised at low potentials. This strain also had a greater capacity to adhere to glass or graphite. This was associated with major changes in lipopolysaccharide structure and exopolysaccharide content. Further comparative genome-scale studies of these strains are expected to yield additional insights into the features that contribute to enhanced extracellular electron transfer.

These results have obvious implications for enhancing the power output of microbial fuel cells. Furthermore, these findings demonstrate that cultures that appear to be "pure", even with the most advanced sequencing techniques, may

in fact contain variants, which have substantial differences in genomic content and physiological capabilities, that can only be detected by forcing the culture through an extreme bottleneck that strongly favors the growth of the rare variant. This phenomenon may be an important consideration for the interpretation of other adaptation studies.

GTL

Bioinformatic Analysis of Gene Regulation in *Geobacter sulfurreducens:* an Integration of Transcriptome and Sequence Information, Molecular Evolutionary Studies, and Database Management

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

[1]University of Tennessee Health Science Center, Memphis, Tenn.; [2]University of Massachusetts, Amherst, Mass.; and [3]Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, México

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

Understanding regulation of gene expression in the deltaproteobacterial family *Geobacteraceae* is critical for our ability to gain insight into the cellular processes which allow these bacteria to participate in environmental bioremediation and energy production. To date, a significant amount of experimental data has been accumulated, providing information about the genome structure of these organisms and describing the changes in their expression at the transcriptome and the proteome level. We have been employing bioinformatic approaches to identify regulatory interactions by synthesizing available knowledge on transcriptional regulators and their predicted and/or experimentally identified target genome sites gained from these experimental studies.

To address the need for a systematic, comprehensive resource integrating available regulatory information, we have developed an online database, GSEL (G*eobacter* Sequence Elements), which compiles regulatory information for *Geobacter sulfurreducens,* an intensively studied model representative of *Geobacteraceae* whose regulation is currently best understood. We have recently completed the development of a new, significantly expanded and updated, relational version 2 of the GSEL database and its accompanying online query system, which compiles manually curated information on operon organization and transcription regulatory elements in the genome of *G. sulfurreducens*. GSEL v. 2 incorporates a new online graphical browser, and it provides significantly expanded search capabilities allowing users to query the database to identify and view graphically the operon structure and regulatory sequence elements in a genome region of interest. Users can search the database by providing genome coordinates, operon ID, or gene ID. Users can also identify transcriptional regulatory sites recognized by a specific transcriptional regulator, those generated in a specific microarray experiment, or those predicted using a specific search method. This new version of the GSEL database includes updated operon predictions along with new information on predicted and/or experimentally validated genome regulatory sites, including promoters, transcriptional factor binding sites, transcriptional attenuators, ribosome-binding sites, and terminator sequences. New features in GSEL v. 2 also include links to information from microarray experiments stored in public gene expression databases (ArrayExpress and Gene Expression Omnibus), and to original publications in PubMed or in other bibliography resources describing how particular regulatory interactions were identified. The development of the relational GSEL database version 2 and its online query system was performed using the tools of LAMP (Linux, Apache, MySQL, and PHP) web development, along with components developed using Java, XML, and JavaScript. The GSEL database version 2, its accompanying online search system and a graphic genome browser provide a unique and comprehensive tool cataloguing information about transcription regulation in *G. sulfurreducens*, which aids in the investigation of mechanisms that regulate its ability to generate electric power, bioremediate environmental waste, and adapt to environmental changes.

In addition to continuing development of database resources, we are also continuing our studies to understand regulatory processes which affect the ability of *Geobacteraceae* to participate in energy production. In our earlier studies, we investigated in detail the target genes and promoters regulated by RpoN, an alternative RNA polymerase sigma factor, which regulates a variety of important cellular processes in *G. sulfurreducens.* We indentified multiple binding sites for RpoN upstream of genes encoding components or regulators of flagellar biosynthesis, ion transport, nitrogen metabolism, signal transduction, multiple *c*-type cytochromes, members of ABC-type branched-chain transporter system, and other important processes and cell systems.

The most highly conserved RpoN-regulated promoter was located upstream of an operon containing a gene for flagellar transcription factor FliA (σ^{28}) and other genes for flagellar biosynthesis. Combined with the absence of the master regulator FlhCD in *Geobacter sulfurreducens,* these results strongly indicate the importance of the *Geobacter sulfurreducens* RpoN sigma factor in flagellation.

Our current studies are focusing on an investigation of the roles of several transcription regulatory systems that are likely to be involved in RpoN-dependent regulatory pathways. In particular, we have investigated the target regulatory sites for an enhancer binding protein, PilR, which participates in RpoN-dependent transcriptional regulation of the *pilA* gene encoding structural pilin. We have predicted multiple PilR-regulated sites upstream of multiple operons related to biosynthesis, assembly, and function of pili and flagella, type II secretory pathways, and cell wall biogenesis. A number of these sites have been found to co-occur with RpoN-regulated promoters, providing a further insight into the regulatory role of RpoN in members of the PilR regulon. We have also been investigating sequence changes and molecular classification of the TetR family of transcriptional regulators in *Geobacteraceae*. In *Geobacter sulfurreducens,* we identified RpoN-regulated promoters upstream of several operons containing genes encoding TetR family regulators. The genome of *G. sulfurreducens* contains nine genes encoding TetR family members. Some of these genes are located immediately upstream of operons encoding functionally important *c*-type cytochromes, e.g., *omcB* and *omcC*, which have been suggested to be actively involved in iron reduction. In order to better understand the role played by the TetR family of transcriptional regulators in the ability of *Geobacteraceae* to participate in electron transfer, we have undertaken a systematic molecular evolutionary study by inferring phylogenetic relationships of the *Geobacteraceae* TetR proteins to their homologs in other species of bacteria. This has allowed us to identify conserved and variable domains in these *Geobacter* proteins, which may be important for the diversity of their functional roles, and to classify them into subgroups based on their sequence similarities. We are currently focusing on the classification of molecular changes in these proteins that distinguish them from their close homologs. Combined with our earlier bioinformatic research on RpoS- and RpoD-dependent regulation in *Geobacter sulfurreducens,* these studies are allowing us to obtain a systematic, detailed view of specific regulatory components affecting transcriptional regulation of electron transfer pathways in *Geobacteraceae.*

GTL

Dynamic Genome-Scale Modeling of *Geobacter* Species in Subsurface Environments

R. Mahadevan,[2]* Jun Sun,[1] Kai Zhuang,[2] Yilin Fang,[3] Tim Scheibe,[3] Hanno Richter,[4] Carla Risso,[4] Paula Mouser,[4] Ashley Franks,[4] and **Derek Lovley[4]**

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

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

Previous studies of the *in situ* bioremediation of uranium-contaminated groundwater have demonstrated that *Geobacter* species are the dominant members of the groundwater community during active bioremediation and the primary organisms catalyzing U(VI) reduction. The addition of acetate to the subsurface to stimulate U(VI) reduction has been shown to promote the growth of *Geobacter* species. Fe(III) serves as the primary electron acceptor and U(VI) is simultaneously reduced. However, the dynamics of growth and activity of *Geobacter* species and their interactions with other environmental factors such as groundwater flow and geochemistry are not well understood. Furthermore, the *Geobacter* species cooperate and compete with other subsurface microorganisms and these interactions are also expected to influence the effectiveness of bioremediation. Hence, the objectives of this program were to develop comprehensive metabolic models of *Geobacter* species and other environmentally relevant microorganisms and to integrate these models with descriptions of other physico-chemical phenomena in order to better understand and optimize bioremediation strategies.

For example, recent molecular ecology studies have demonstrated that *Geobacter* speices must compete with acetate-oxidizing, Fe(III)-reducing *Rhodoferax* species in the subsurface. Prior to the addition of acetate to stimulate U(VI)

reduction, *Rhodoferax* competes well with *Geobacter* species at the uranium-contaminated field study site in Rifle, CO. In fact, in zones where substantial quantities of ammonium are available, *Rhodoferax* speices are more prevalent than *Geobacter* species. In order to expand our limited knowledge about Fe(III)-reducing *Rhodoferax* species, *R. ferrireducens* was studied with a multi-faceted approach that included the description and analysis of the genome annotation, investigation of new metabolic capabilities, and the development of a constraint-based *in silico* metabolic model. The model provided new insights into the stoichiometry of the electron transport chain and the efficiency of substrate utilization under different conditions.

The newly developed *R. ferrireducens* genome-scale model was coupled with our previously developed *G. sulfurreducens* model in order to understand the competition between these organisms before and during *in situ* bioremediation of uranium-contaminated subsurface environments. This also required characterization of the acetate transport systems and their uptake kinetics. The simulation of the competition between these organisms suggested that the competition between them is modulated by two factors: the ability of *G. sulfurreducens* to fix nitrogen under ammonium limitation, and a rate vs. yield trade-off between these two organisms. Prior to acetate amendment, if ammonium is limited, *G. sulfurreducens* dominates due to its ability to fix nitrogen. However, if the system contains abundant ammonium, *R. ferrireducens*, which has higher biomass yields, is favoured because the acetate flux is very low. During acetate amendment, the high acetate flux strongly favours *G. sulfurreducens*, a rate-strategist. The model also predicts high respiration rates by *G. sulfurreducens* during nitrogen fixation at the expense of biomass yield, leading to an increase in U(VI) reduction under low ammonium conditions. The results of the simulation agreed well with the subsurface community composition determined from molecular analysis of subsurface samples before and during *in situ* bioremediation at the Rifle site.

In an attempt to better understand and predict how *Geobacter* species might grow and metabolize in the subsurface under diverse conditions that might be imposed in different bioremediation strategies the genome-scale, constraint-based model of the metabolism of *G. sulfurreducens* was coupled with the reactive transport model HYDROGEOCHEM. The initial modeling was simplified by only considering the influence of three growth factors: acetate, the electron donor added to stimulate U(VI) reduction; Fe(III), the electron acceptor primarily supporting growth of *Geobacter*; and ammonium, a key nutrient. The constraint-based model predicted that growth yields of *Geobacter* varied significantly based on the availability of these three growth factors and that there are minimum thresholds of acetate and Fe(III) below which growth and activity are not possible. This contrasts with typical, empirical microbial models which assume fixed growth yields and the possibility for complete metabolism of the substrates. The coupled genome-scale and reactive transport model predicted acetate concentrations and U(VI) reduction rates in a field trial of *in situ* uranium bioremediation that were comparable to the predictions of a calibrated conventional model, but without the need for empirical calibration required for conventional modeling, other than specifying the initial biomass of *Geobacter*. These results suggest that coupling genome-scale metabolic models with reactive transport models may be a good approach to developing models that can be truly predictive, without empirical calibration, for evaluating the probable response of subsurface microorganisms to possible bioremediation approaches prior to implementation.

One of the limitations of the acetate-amendment strategy is that over time, Fe(III) oxides are depleted near the site of acetate injection and acetate-oxidizing sulfate reducers, which are ineffective in U(VI) reduction, become predominant, consume the acetate, and the effectiveness of the bioremedation deteriorates. It has been proposed that this problem might be circumvented by adding lactate instead of acetate with the hope that lactate would stimulate the growth of the lactate-oxidzing sulfate-reducing *Desulfovibrio* species which are highly effective U(VI) reducers. Furthermore, *Desulfovibrio* species only incompletely oxidize lactate to acetate, and thus acetate generated from the metabolism of lactate could potentially support the growth of *Geobacter* species, but the growth of acetate-oxidizing sulfate reducers would be prevented because *Desulfovibrio* would have consumed the sulfate. Before incurring the cost and time of implementing this new bioremediation strategy it is desirable to evaluate its likely outcome *in silico*. Therefore, we developed a computational model of *Geobacter*, *Desulfovibrio*, and the acetate-oxidizing sulfate-reducing bacterial community by dynamically integrating the genome-scale metabolic model of *G. sulfurreducens*, a central metabolic model of *D. vulgaris,* and the kinetic model of a representative acetate-oxidizing sulfate reducer using the dynamic multi-species metabolic modeling framework. Simulations of batch, continuous, and fed-batch lactate injection indicated that the syntrophic community based on acetate exchange between *D. vulgaris* and *G. sulfurreducens* can lead to prolonged Fe(III) reduction and can potentially accelerate U(VI) reduction. The next step will be to couple this model with the reactive transport model to design optimal feed strategies for this bioremediation approach.

These studies demonstrate that genome-scale modeling can aid in the interpretation and prediction of microbe-microbe and microbe-geochemical interactions in the subsurface. In addition to enhancing the understanding of basic subsurface microbiology, this approach has clear practical application for bioremediation optimization.

GTL

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

Barbara Methé[1]* (bmethe@jcvi.org), Robert Deboy,[1] Qinghu Ren,[1] Michael Geimer,[1] Jamison McCorrison,[1] Dawn Holmes,[2] and **Derek Lovley**[2]

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

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

The Old Rifle site located in Rifle, Colorado represents a former uranium ore-processing facility which is currently being managed as a part of the Uranium Mill Tailings Remedial Action (UMTRA) program of the U.S. Department of Energy. Over time, continuous leaching of radionuclides from processed mill tailings present at this location resulted in contamination of both the groundwater and sediment of the surrounding aquifer. Currently, in situ bioremediation pilot studies, specifically the biostimulation of members of the *Geobacteraceae* through the addition of acetate to groundwater, is being investigated at this site as a possible corrective strategy. The study presented here is meant to complement these wider ongoing investigations through the application of next generation sequencing and metagenomic analyses in an effort to reveal genetic information from subsurface microorganisms without the need for prior cultivation.

A metagenomic framework for use in this project has been established as follows. Genomic DNA from the 2007 sampling of the Rifle D05 site groundwater was used to prepare a 3-4 kb insert paired-end library for sequencing using 454 FLX pyrosequencing. One run of 454 sequencing has been completed yielding ~35% paired-end reads and 65% fragments (single reads) which generated 127,849,360 total bases of an average read length of 245bp and an average quality value of 33.

Initial analyses of this data have focused on the evaluation of a variety of computational methods to investigate taxonomic and functional classifications of both the unassembled and assembled reads. Assemblies were completed using the Newbler Assembler from 454 and a modified version of the Celera Assembler (CA). A greater N50 contig and scaffold size, respectively (N50 contig= 2419 vs. 1339 and N50 scaffold= 19098 vs.15586) was achieved with CA. In addition, results of whole genome alignments using Nucmer demonstrated greater coverage of reference genomes from the *Geobacteraceae* with the CA versus the Newbler data set. Therefore, the CA assembly has been used for additional analyses of assembled data.

Taxonomic classification of individual reads and contigs was completed using the Metagenome Analyzer (MEGAN) program. Results from this analysis identified ~57% of all reads assigned at the Family level were classified as members of the *Geobacteraceae* (*Geobacteraceae* + *Pelobacteraceae* based on NCBI taxonomy). From an examination of assembled contigs, approximately 55% were assigned to the genus *Geobacter* with members of the Subsurface Clade 1 including *G.* sp. M21, *G. bemidjiensis* and *G. uraniumreducens* representing the predominate species classifications.

Functional classifications of the individual 454 reads were performed using the MG-RAST pipeline. In this analysis a total of 499,050 reads representing ~121 Mb were examined of which 195,949 sequences (39%) could be matched to proteins in SEED subsystems using an e-value cut-off of $1e^{-5}$. The top five functional categories determined were: Carbohydrate Metabolism (11.6%), Clustering-based Subsystems (10.6%), Virulence (9.4%), Amino Acids and Derivatives (7.5%) and Protein Metabolism (7.1%). Carbohydrate Metabolism was dominated by reads placed in the subcategory of central intermediary metabolism (~24% of the total for the category) but reads related to functions such as biofilm formation and, methanogenesis and one-carbon metabolism, were also found. The category of Virulence was dominated by reads classified in the subcategories of resistance to antibiotics and toxic compounds (36%), and prophage and transposons (29%). However, classification of reads in subcategories such as Type IV pilin as well as Type III and Type VI secretion systems which are often found in pathogenic organisms, but more recently have been identified in non-pathogenic bacteria, were also noted.

From the assembled data, the program, Metagene, was used to predict ORFs from the contigs generated by CA. A total of 8105 ORFs consisting of 4274 complete and 3831 partial ORFs were predicted. Analysis of annotation of these ORFs has suggested that even though the majority of assembled genomic data is related to members of the *Geobacteraceae*, many ORFs (at least one-third) identified are not closely related to sequenced organisms in culture. These results suggest that even among the *Geobacteraceae*, a greater diversity of organisms exists at this site than has currently been captured in pure culture and sequenced.

In addition to examining the D05 microbial community at the level of the genome, we have also begun experimentation designed to elucidate community gene expression through investigation of the metatranscriptome. We have used a capture oligonucleotide approach to enrich mRNA through the removal of 16S or 23S rRNA molecules. The enriched mRNA is then used as template in cDNA reactions which can subsequently be sequenced via 454 pyrosequencing. Currently we have produced cDNA from enriched mRNA and have tested the efficacy of the cDNA through a series of RT-PCR-based experiments targeting a set of housekeeping genes that are generally constitutively expressed. Included in this series of validations has been the design of primer sequences to these housekeeping targets based on the metagenomic sequence. From D05 cDNA we have used these primers to amplify a region of the target gene and validated these results by sequencing and analysis of the gene product.

Additional investigations are now underway to further examine the D05 metagenome. These include the use of: 1) an iterative protein clustering approach to further identify and characterize protein families from individual reads and predicted ORFs, and 2) the application of a web based tool developed at the JCVI, Advanced Reference Viewer, designed to facilitate comparative analysis of metagenomic datasets in the context of a reference genome and 3) comparison of the metatranscriptome (gene expression) to the metagenome. Overall, the efforts presented here are central to several goals of this project. Since the samples examined were obtained from bulk community DNA and RNA as opposed to individual populations, the resulting information may serve to generate new insights and hypotheses concerning metabolic processes and interactions between community members. Further, these results may also serve as a framework for predictive modeling of processes relevant to bioremediation.

GTL

New Insights into *Rhodoferax ferrireducens* Through Genome Annotation and Genome-Based *in silico* Metabolic Modeling

Carla Risso[1]* (crisso@microbio.umass.edu), Jun Sun,[2] Kai Zhuang,[3] Radhakrishnan Mahadevan,[3] Olivia Bui,[2] Robert DeBoy,[4] Wael Ismael,[1] Christophe H. Schilling,[2] Barbara A. Methé,[4] and **Derek Lovley**[1] (dlovley@microbio.umass.edu)

[1]Dept. of Microbiology, University of Massachusetts, Amherst, Mass.; [2]Genomatica Inc., San Diego, Calif.; [3]University of Toronto, Canada; and [4]J. Craig Venter Institute, Rockville, Md.

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

The Fe(III)-reducing bacterium *Rhodoferax ferrireducens* is a metabolically versatile organism known for its ability to convert sugars to electricity and its important role in dissimilatory metal reduction in some subsurface environments. In order to expand our limited knowledge about *R. ferrireducens*, a multi-approach study was conducted, including the description and analysis of the genome annotation, the development of a constraint-based genome-scale metabolic model, and physiological studies.

The *R. ferrireducens* genome consists of a circular chromosome of 4,712,337 base pairs and a plasmid with 257,447 bp. A total of 4,770 protein-encoding open reading frames (ORFs) have been predicted in the chromosome, of which 2,797 have been assigned a putative function. Analysis of the top BLASTP hits of *R. ferrireducens* ORFs against a database of complete genomes established the following taxonomic breakdown: *Betaproteobacteria* (3,253 genes), *Gammaproteobacteria* (258), *Alphaproteobacteria* (185), *Delta/Epsilonproteobacteria* (121). Of completed genomes, those representing *Polaromonas sp.* are the most closely related to *R. ferrireducens*, followed by *Acidovorax sp.*, *Verminephrobacter eiseniae*, *Methylibium petroleiphylum* and *Ralstonia* sp. There are 169 *R. ferrireducens* genes whose best match is to another *R. ferrireducens* gene suggesting these might be lineage-specific duplications. Conversely, about 14% of the *R. ferrireducens* genes are hypothetical ORFs.

The genome annotation of *R. ferrireducens* was used for the construction of the metabolic network. The annotated genes of the *R. ferrireducens* genome, as well as genes from several high-quality genome-scale metabolic models, were utilized to generate a draft network as a starting point for model reconstruction. The draft model successfully captured significant portions of central metabolism, as well as biosynthetic pathways for amino acids, nucleotides, and lipids. Among the base models used, *E. coli* was the phylogenetically closest to *R. ferrireducens* and provided about half of all reactions in the draft model. Of the 4770 genes in the *R. ferrireducens* genome, 744 genes were included in the reconstructed genome-scale network. The *R. ferrireducens* metabolic model contains 762 reactions and 790 metabolites including 69 extracellular metabolites. The metabolic capabilities of the *R. ferrireducens* network were calculated using flux balance analysis and linear optimization. Biomass synthesis was

* Presenting author

selected as the objective function to be maximized in growth simulations, and ATP consumption was selected as the objective function to be maximized in energy requirement simulations.

The *in silico* model was used to estimate energy parameters. An optimization algorithm iterated the $H^+/2e^-$ ratio of NADH dehydrogenase from 1 to 4, the $H^+/2e^-$ ratio of cytochrome reductase from 1 to 4, and non-growth associated maintenance (nGAM) from 0 to 2.5 mmol ATP/gDW/hr to estimate the values that would result in the most consistent growth yields between those experimentally determined versus those predicted by the model. This process identified optimal energy parameters of an $H^+/2e^-$ ratio of 2 for both NADH dehydrogenase and cytochrome reductase, and an nGAM of 0.45 mmol ATP/gDW/hr. This set of energy parameters was applied to the model and validated by the comparison between *in silico* predictions and experimentally determined yields.

Several previously unknown metabolic capabilities of *R. ferrireducens* were discovered. These included the ability to utilize carbohydrates other than glucose to support growth, including fructose and mannose. The catabolism of certain disaccharides was also determined. Notably, cellobiose degradation poses biotechnological interest because of the potential of turning common cellulosic waste products into energy. It was found that *R. ferrireducens* is capable of fumarate dismutation, where fumarate is used as electron donor, electron acceptor and carbon source. This metabolism was analyzed by model simulations and the calculated yields closely matched the actual experimental results. Examination of the genome suggested that *R. ferrireducens* was able to utilize citrate. A gene encoding CitT is 44% identical to a citrate transporter from *E. coli* and is located in a cluster of genes also associated with citrate metabolism. Experimental work confirmed that *R. ferrireducens* can indeed grow on citrate as electron and carbon source with either Fe(III) or nitrate as electron acceptor. The *in silico* model predicted that *R. ferrireducens* could completely oxidize citrate with Fe(III) as electron acceptor and suggested four other pathways of citrate oxidation that also produce acetate and/or succinate under electron acceptor limiting conditions. No fermentative growth was observed in *R. ferrireducens*. Fermentative growth on glucose was simulated with the *in silico* model described above and the result confirmed the experimental observation. Detailed analysis of the metabolic network suggested that the inability of fermentative growth of *R. ferrireducens* is likely due to 1) an inability to convert acetyl-CoA to ethanol, which is in turn consistent with its inability to utilize ethanol; 2) the requirement of ubiquinol in synthesis of biomass components and NADH dehydrogenase as the only carrier able to transfer electrons from NADH to ubiquinone. Thus, a terminal electron acceptor is required for cell growth.

The model proved to be a useful aid in determining which substrates might best support growth for specific applications. For example, growth on eight representative electron donors was simulated under donor-limiting conditions with Fe(III) as electron acceptor. Acetate and glycolate resulted in the lowest biomass yield, whereas cellobiose produced the highest. Glucose yielded higher biomass than citrate or benzoate.

Genome analyses further suggested that *R. ferrireducens* possesses strategies to manage a number of environmental challenges including aromatic compound metabolism and stresses due to heavy metals and oxidation. *R. ferrireducens* can also cope with nutrient limitation by synthesizing polyhydroxyalkanoates. These storage molecules have industrial and medical interest due to their properties as thermoplastics and elastomers.

This multi-approach study has provided new insights into the remarkable metabolic versatility of *R. ferrireducens* and its potential practical applications. The genome scale model of *R. ferrireducens* will be an important tool for further analysis of the ecology of anaerobic subsurface environments and for the optimization of the unique current-producing capabilities of this organism.

GTL

Adaptive Evolution of *Geobacter sulfurreducens* under Likely Subsurface Bioremediation Conditions Revealed with Genome Resequencing

Zarath M. Summers[1]* (zsummers@microbio.umass.edu), Pier-Luc Tremblay,[1] Toshiyuki Ueki,[1] Wael Ismail,[1] Jennifer K. Hart,[1] Bernhard Palsson,[2] Christian Barrett,[2] Karsten Zengler,[2] Yuan Gao,[3] Bin Xie,[3] Chen Liang,[3] and **Derek R. Lovley**[1]

[1]Dept. of Microbiology, University of Massachusetts, Amherst, Mass.; [2]University of California San Diego, La Jolla, Calif.; and [3]Virginia Commonwealth University, Richmond, Va.

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

In situ bioremediation of subsurface environments contaminated with organic and/or metal pollutants typically relies on the activity of members of the natural subsurface microbial community. The conditions imposed on that community in order to optimize bioremediation can present circumstances that the subsurface organisms have never previously experienced. Therefore, there may be substantial selective pressure on subsurface microorganisms to change their physiological properties by accumulating beneficial mutations during the implementation of subsurface bioremediation. If so, this could result in dramatic improvements in bioremediation rates and/or effectiveness over time. Thus, documenting which mutations accumulate could aid in better understanding the physiology and ecology of complex bioremediation reactions as well as the interactions of the bioremediating microorganisms with their environment.

For example, *Geobacter* species have been shown to play an important role in the degradation of a diversity of organic contaminants in subsurface environments and are the predominant microorganisms during removal of uranium and vanadium from contaminated groundwater during in situ bioremediation of these contaminants with added acetate. The growth of *Geobacter* species in uncontaminated subsurface environments is generally limited by low availability of electron donors and the difficulties in accessing insoluble Fe(III) oxides, their primary electron acceptor. Therefore, there is little selective pressure for rapid metabolism and growth. This contrasts with the abundance of electron donor when organics are added to the subsurface to stimulate dissimilatory metal reduction or that result from contamination of the subsurface with organic compounds. Furthermore, some of the organic compounds that may be added to the subsurface in some bioremediation strategies are not normally important electron donors in the subsurface and their sudden availability can significantly change the rules of substrate competition.

As an initial evaluation of the potential for adaptive evolution during bioremediation involving *Geobacter* species several studies simulating selective pressures that might be imposed during bioremediation were conducted with *Geobacter sulfurreducens*. For example, slow-release lacatate polymers represent a simple method of providing electron donor in the subsurface that may be preferable to the current practice of continually pumping acetate into the subsurface. *G. sulfurreducens* grows poorly on lactate, but the genome-scale metabolic model for this organism predicted that it could potentially grow faster on lactate than on its common electron donor, acetate. Five parallel strains of *G. sulfurreducens* were continually transferred in medium with lactate as the sole electron donor and carbon source. Over a two-year period the doubling time of the lactate cultures decreased from 22 hours to 5 hours, yielding strains that could grow as fast on lactate as on as the ancestral culture grew on acetate.

The lactate-adapted strains were examined with a combination of resequencing chip technology and Illumina sequencing of genomic DNA. All of the lactate-adapted strains had single-base pair substitutions in the gene, GSU0514. Further investigation, including DNA binding and footprint assays, revealed that GSU0514 encodes a transcription factor that, when expressed, represses transcription of succinyl-CoA synthetase. Succinyl-CoA synthetase is an important TCA cycle enzyme that converts succinyl-CoA to succinate with the generation of ATP. Succinyl-CoA synthetase activity is not required for growth on acetate because succinyl-CoA can be converted to succcinate by acetylCoA-transferase, which simultaneously activates acetate to acetyl-CoA for oxidation in the TCA cycle. However, during growth on lactate, there is not sufficient free acetate to accept CoA from succinyl-CoA and thus succinyl-CoA synthetase activity is required. Microarray analysis of gene expression in one of the five adapted strains revealed that expression of the succinyl-CoA synthetase was upregulated with transcript abundance 6.5 fold higher than in wild-type cells. Furthermore, succinyl-CoA synthetase enzymatic activity could not be detected in wild-type cells but was present in the adapted strains.

These results suggested that the mutations in GSU0514 limited the binding capacity of this transcriptional regulator. When GSU0514 was deleted from wild-type cells the knock-out mutant had enhanced growth on lactate, but slightly slower than the adapted strain. Knocking in a copy of GSU0514 that had the same single base-pair change found in one of the adapted strains resulted in a wild-type strain that grew as well on lactate as the adapted strain. The difference in growth on lactate between removing the gene and slightly altering its sequence suggests that this transcription factor influences the expression of a diversity of genes. This is consistent with the finding that over 100 additional genes are differentially expressed in the adapted strain at more than a two-fold difference, including response regulators.

The greater availability of electron donors during bioremediation, alleviates the limitation of growth by electron donors, establishing selective pressure for strains of *Geobacter* that can utilize Fe(III) oxides more rapidly. Over a two-year period *G. sulfurreducens* was continually transferred as rapidly as possible in medium with Fe(III) oxide as the sole electron acceptor and unlimited electron donor availability. The adapted strains now reduce the same amount of Fe(III) in seven days that the ancestral, wild-type strain took 40 days to reduce. Resequencing the genome of one of the adapted strains revealed a single base pair change in the regulatory region 99 bases up stream of the gene that encodes the outer-membrane, c-cytochrome, OmcR. Although microarray data is not yet available, quantitative PCR analysis demonstrated that transcripts of omcR were 19-fold more abundant in the adapted strain than in ancestral strain. These results are of interest because they implicate OmcR in Fe(III) oxide reduction. The appropriate knock-out and knock-in mutations in omcR and the putative regulatory region are currently under construction.

These studies demonstrate that single base pair mutations in regulatory regions of the *G. sulfurreducens* genome that accumulate as the result of new selective pressures that *Geobacter* species are likely to experience during in situ bioremediation

of subsurface environments can dramatically impact on the range of electron donors utilized and the capacity for Fe(III) oxide reduction. Thus, long-term bioremediation of subsurface environments may be viewed as large-scale adaptive evolution experiments and the possibility for improvements in microbial performance as a consequence of this evolution should be considered when attempting to predict the outcome and performance of different bioremediation strategies.

GTL

Novel Mechanisms Regulating the Expression of Genes in *Geobacter* species Important for Metal Reduction, Electricity Production, and Growth in the Subsurface

Toshiyuki Ueki* (tueki@microbio.umass.edu), Byoung-Chan Kim, and **Derek R. Lovley** (dlovley@microbio.umass.edu)

Dept. of Microbiology, University of Massachusetts, Amherst, Mass.

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

Genome-scale *in silico* models of *Geobacter* species are currently being developed to predict the activities of *Geobacter* species during *in situ* uranium bioremediation and electricity-harvesting from waste organic matter and renewable biomass. To implement regulatory modules in these models, we have been investigating molecular mechanisms regulating gene expression in *Geobacter* species under various environmentally relevant conditions.

For example, the genomes of *Geobacter* species encode an unusually large number of homologues of two-component systems. This may reflect the need for *Geobacter* species to adapt to a wide range of conditions in subsurface environments. We discovered a novel two-component system in *G. sulfurreducens*, designated GsuTCS1. The sensor histidine kinase component has a novel sensor domain, which contains two putative *c*-type heme binding motifs. The response regulator is in the RpoN-dependent enhancer-binding protein family. Our previous studies suggested that GsuTCS1 was involved in redox sensing, Fe(III) reduction, and biofilm formation. Biofilm formation is an important process for optimal current production in microbial fuel cells. To better understand the function of GsuTCS1, the transcriptome of a strain that overproduces GsuTCS1 was compared with the transcriptome in the wild-type cells. Genes with higher transcript abundance in the overproducing strain encode membrane proteins such as a cytochrome *c*, a transporter, and a porin. Biochemical analyses further identified direct target genes for GsuTCS1, most of which encode hypothetical proteins with unknown function, but these hypothetical proteins were predicted to have a signal peptide, suggesting that they function in the cellular membrane. These results suggest that GsuTCS1 regulates genes involved in modifying membrane structures during redox sensing, Fe(III) reduction, and electricity production.

Geobacter species have been shown to fix atmospheric nitrogen during *in situ* uranium bioremediation and this ability seems to be a key to the predominance of *Geobacter* species in a variety of subsurface environments. We have identified novel regulatory cascades controlling gene expression during nitrogen fixation in *Geobacter* species by a systematic analysis integrated by functional and comparative genomics in combination with biochemical and genetic methods. Unlike regulatory mechanisms known in other nitrogen-fixing microorganisms, the nitrogen-fixation gene regulation in *Geobacter* species was controlled by two two-component His-Asp phosphorelay systems. One of the systems, GnfL/GnfM, was the master regulator that activated transcription of the majority of nitrogen-fixation genes. In addition, the GnfL/GnfM system repressed the gene encoding glutamate dehydrogenase during nitrogen fixation. The GnfL/GnfM system appeared to be essential for growth even in the presence of fixed nitrogen, as a deletion mutant of the GnfL/GnfM system could not be obtained. Overexpression of the GnfL/GnfM system resulted in induction of the nitrogen-fixation genes and repression of the glutamate dehydrogenase gene in the presence of fixed nitrogen, suggesting that the GnfL/GnfM system regulates the nitrogen-fixation genes as well as the glutamate dehydrogenase gene. The other system, GnfK/GnfR, which was directly activated by the GnfL/GnfM system, was shown to be essential for growth in the absence of fixed nitrogen. The amino acid sequence analysis of GnfR indicated that GnfR was an antiterminator. Deletion of *gnfK* and *gnfR* resulted in premature transcription termination of a subset of the nitrogen-fixation genes, whose transcription is activated by the GnfL/GnfM system and which have transcription termination signals in their promoter regions. These results further demonstrate that the GnfK/GnfR system controls by transcription antiterminiton the expression of the subset of the nitrogen-fixation genes. This study provides a new paradigm to nitrogen-fixation gene regulation.

The TCA cycle is important in *Geobacter* species because it represents the main pathway for the generation of energy and serves to synthesize precursor metabolites. For instance, *in situ* transcript levels of the gene for the citrate synthase, a key enzyme in the TCA cycle, can serve as a biomarker to monitor the metabolic activities of *Geobacter* species during uranium bioremediation and electricity-harvesting. Therefore, the regulation of genes involved in biosynthesis and energy generation was investigated. We identified a transcriptional repressor, designated HgtR, which regulates the expression of genes involved in biosynthesis and energy generation. HgtR is a novel transcription factor, whose homologues are only found in *Geobacter* species. The expression of the *hgtR* gene increased during growth with hydrogen as the electron donor. A strain in which *hgtR* was deleted could not grow on hydrogen. The deletion or overexpression of *hgtR* resulted in activation or repression, respectively, of genes involved in biosynthesis and energy generation. These results suggest that HgtR is a global transcriptional repressor that regulates the genes involved in biosynthesis and energy generation in *Geobacter* species. Moreover, it appeared likely that *hgtR* expression was regulated by the enhancer-binding protein containing a hydrogenase domain. This study paves the way to better understanding of genome-scale metabolic gene regulation in *Geobacter* species.

Riboswitches, noncoding RNA elements found in the untranslated region of mRNA, sense and bind cellular metabolites to control gene expression. *Geobacter* species have been predicted to contain riboswitches sensing a variety of cellular metabolites. For instance, cyclic di-GMP, which functions as a second messenger to regulate diverse physiological processes in bacteria, might be sensed by riboswitces in *Geobacter* species, as homologues of the riboswitch that is known to sense cyclic di-GMP in other bacteria were identified in the genomes of *Geobacter* species. *G. uraniireducens* has the largest number of cyclic di-GMP riboswitch homologues among bacteria whose genomes have been sequenced. In *G. sulfurreducens*, genes known to be differentially regulated during metal reduction and electricity production, such as *omcS* and *omcT*, were found to contain a cyclic di-GMP riboswitch signature in their noncoding region of mRNA. The 5' untranslated region of *omcS* mRNA was shown to be critical for *omcS* expression. These suggest that riboswitches play a role in the regulation of key genes involved in metal reduction and current production.

This increased understanding of regulation of process central to the growth of *Geobacter* species in subsurface environments and the optimization of groundwater bioremediation and electricity production not only contributes to a better understanding of the physiology and ecology of *Geobacter* species, but also provides important information for developing new bioremediation and electricity harvesting strategies.

Nitrogen Metabolism in the *Geobacteraceae*: The Role of RpoN

Karsten Zengler[1]* (kzengler@ucsd.edu), Yu Qiu,[1] Jessica De Ingeniis,[2] **Derek Lovley**[3] (dlovley@microbio.umass.edu), and Bernhard Ø. Palsson[1]

[1]Dept. of Bioengineering, University of California San Diego, La Jolla, Calif.; [2]Burnham Institute for Medical Research, La Jolla, Calif.; and [3]Dept. of Microbiology, University of Massachusetts, Amherst, Mass.

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

Nitrogen assimilation is essential for the production of proteins, nucleic acids and cell wall components in prokaryotes and is therefore under stringent control. To study nitrogen assimilation and its control we used genome-wide binding profiles of RNA polymerase (RNAP) and major sigma factors σ^{70} and σ^{N} in *G. sulfurreducens* under various conditions by ChIP-chip analysis in combination with gene expression profiles using high-density tiling arrays. Over 150 binding sites for σ^{N} (RpoN) were experimentally identified in *G. sulfurreducens.* All key genes for N-assimilation, nitrogenase, glutamate dehydrogenase, glutamine synthetase and glutamate synthase, were controlled by RpoN, proofing that RpoN is the global regulator for N-metabolism in *G. sulfurreducens*. Besides the important role in controlling nitrogen metabolism, RpoN was identified to be an essential sigma factor controlling various other aspects of metabolism in *G. sulfurreducens.* For examples genes responsible for ferrous iron transport were under σ^{N} (and Fur) control. These genes were only expressed under fumarate reducing conditions in fast growing cells (NH_3 assimilating) but not in slow growing cells (N_2 assimilating). These genes were also not expressed under Fe(III)-reducing conditions (slow growing cells).

Besides ammonium and nitrogen gas we determined that *G. sulfurreducens*, *G. uraniireducens* and *G. metallireducens*

were all able to assimilate glutamine as sole N-source although cells grow significantly slower with glutamine as N-source. Gene expression profiles of *G. sulfurreducens* demonstrated that a variety of genes (e.g. *nif* genes) were up-regulated while growing with N_2 but also with glutamine as compared to NH_3. Key genes involved in nitrogen assimilation, glutamate dehydrogenase (GDH), as well as glutamine synthetase (GS) and glutamate synthase (GOGAT) were differentially expressed during growth with different nitrogen sources. Cells grown with ammonium as nitrogen source had a higher expression level of GDH while levels of GS and GOGAT were significantly down-regulated compared to growth with N_2 and glutamine, reflecting gene expression profiles for excess of nitrogen and N-limiting condition.

Geobacter strains responded to nitrogen limitation in a similar way as described for *Escherichia coli* and *Salmonella typhimurium*. However, in these bacteria, N-limitation is being sensed by the internal glutamine pool (a high glu/gln ratio represents N-limiting conditions) therefore GS and GOGAT expression is up-regulated. *G. sulfurreducens* did respond to N-limitation with up-regulation of GS and GOGAT but the organism did not perceive external N-limitation as internal glutamine limitation. The intracellular glu/gln ratio was in fact an order of magnitude lower for glutamine grown cells as for cells grown with ammonium. In addition, expression of both genes (GlnE and GlnD) that negatively control GS activity directly by adenylylation and uridylylation was significantly down-regulated in glutamine grown cells. We therefore conclude that members of the *Geobacteraceae* perceive N-limitation differently as shown before for other microorganisms. This finding is in agreement with the proteinaceous sparse environments these organisms are typically found in.

GTL

Structures and Redox Potentials of Periplasmic Cytochromes from *Geobacter sulfurreducens*

P. Raj Pokkuluri[1]* (rajp@anl.gov), Yuri Y. Londer,[1] Norma E.C. Duke,[1] Miguel Pessanha,[2] Leonor Morgado,[2] Ana P. Fernandes,[2] Carlos A. Salgueiro,[2] and **Marianne Schiffer**[1] (mschiffer@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Argonne, Ill.; and [2]Requimte, CQFB, Dep. Quimica, FCT-UNL, Caparica, Portugal

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

The genome of *Geobacter sulfurreducens* (Gs) encodes a large number of c-type cytochromes that are expected to function in metal reduction pathways. Of the 111 cytochromes predicted, 91 were identified as produced in one or the other growth conditions tested.[1] Multiheme cytochromes have been implicated in the electron transport chain(s) used by the organism in the reduction of soluble or insoluble metal oxides, or transfer of electrons to electrodes. In addition, the cytochromes may be serving as capacitors (electron-storage sinks) that allow the continued electron flow from the inner membrane to the periplasm in the absence of suitable electron acceptors. Such an electron storage capability will allow the *Geobacter* species to move towards zones where electron acceptors are available, as suggested by Nunez and co-workers.[2]

The periplasmic cytochromes from the c_7 family are required for the reduction of Fe(III) and U(VI) by Gs.[3,4] The periplasmic triheme cytochrome c_7 PpcA (encoded by GSU0612) was characterized by Lloyd et al., 2003.[5] We expressed PpcA in *E. coli* [6] and determined its three-dimensional structure at 1.45 Å resolution.[7] Further, we identified several sequences homologous to PpcA in the Gs genome.[7] Four of them are of the same size as PpcA; three others are polymers of c_7-type domains, two of which consist of four domains and one of nine domains that contain a total of 12 and 27 hemes per protein, respectively.

We cloned, expressed, purified, crystallized and determined the structures by X-ray diffraction of all four triheme homologs of PpcA, encoded by GSU0364, GSU0365, GSU1024, and GSU1760. Further we characterized the microscopic reduction potentials of the individual hemes in each protein using NMR and UV-Vis spectroscopic techniques. All the c_7 family cytochromes from Gs have very similar heme core structures. Their overall protein fold is similar with some local structural differences observed. Their structural similarity is highest near heme IV and lowest near heme I. All the proteins have a positively charged surface near heme IV. Such a positively charged surface patch was also observed near the equivalent heme IV in the family of cytochromes c_3. The region around heme IV in cytochromes c_3 was suggested to be the interaction site of their physiological partner, hydrogenase.[8] By analogy, this region around heme IV in cytochromes c_7 may be the interaction site with their physiological reductase. It can also be speculated that all cytochromes c_7 in Gs interact with the same partner near heme IV as the structure and surface charge are most conserved in this part of their structures. The structures and surface charges have prominent differences near heme I between each of the cytochromes c_7. Therefore, we suggest that these cytochromes might interact with different partners near heme I or heme III. We determined that the individual heme reduction potentials are different in each protein resulting in different order of oxidation of the hemes in each molecule. This further suggests their functional differences. Studies of mutant Gs strains carrying knockouts of the cytochromes c_7 by Dr. Laurie DiDonato in Prof. Lovley's group displayed different iron reduction rates implying a unique role for each of the proteins.[3]

Multiheme cytochromes that could form protein "nanowires" were identified in Gs and represent a new type of multiheme cytochromes.[7] The sequences of these (two 12-heme GSU1996, GSU0592 and one 27-heme containing GSU2210) proteins suggest that they are formed by domains homologous to triheme cytochromes c_7. In each domain of the above polymers the heme equivalent to heme IV has His-Met coordination whereas in c_7 all three hemes have bis-His coordination. We have previously determined the structure and measured the macroscopic redox potential of one representative domain (domain C) of a dodecaheme cytochrome (GSU1996).[9] Further, we determined the structures of hexaheme containing domains, AB and CD, and the full four domain protein that could be considered a protein wire. Recently, the microscopic redox properties of the heme groups of domain C were determined using NMR and UV-visible spectroscopies. As observed in cytochromes c_7 family, the hemes in domain C also have different microscopic redox potentials: heme I, -71 mV; heme III, -146 mV; heme IV, -110 mV, which are modulated by strong redox interactions that are dominated by electrostatic effects. The order of oxidation of the hemes is III (His-His coordination) – IV (His-Met) – I (His-His). This result is rather surprising as the His-Met coordinated heme IV does not have the highest potential. It suggests that the polypeptide chain surrounding the hemes plays a dominant role in controlling the individual heme potentials.

This work combines high resolution structural studies and detailed microscopic redox potential characterization of a family of periplasmic triheme proteins and a protein domain. Our results demonstrate that the strong redox interactions between closely packed hemes and their specific interactions with the polypeptide surroundings in structurally similar proteins can extend the range of working heme redox potentials. Such a modulation in heme potentials is likely to have a functional significance in the metabolism of Gs.

References

1. Ding, Y.-H. R. et al. *Biochim. Biophys. Acta Proteins & Proteomics*, 1764, 1198-1206 (2006).
2. Esteve-Núñez, A. et al. *Environ. Microbiol.* 10, 497-505 (2008).
3. DiDonato, L.N., Lovley, D.R. et al, private communication.
4. Shelobolina E.S. et al. *BMC Microbiology*, 7:16 (2007).
5. Lloyd J.R. et al. *Biochem. J.* 369, 153-161 (2003).
6. Londer Y.Y. et al. *Biochim. Biophys. Acta Bioenergetics*, 1554, 202-211 (2002).
7. Pokkuluri P.R. et al. *Biochemistry*, 43, 849-859 (2004).
8. Yahata N. et al. *Biochemistry*, 45, 1653-1662 (2006).
9. Pokkuluri P.R. et al. *Prot. Sci*, 13, 1684-1692 (2004).

Caulobacter Systems Understanding

GTL

Automatic Segmentation and Structural Study of the Bacterial Cell Wall in Cryo-Electron Tomography

Farshid Moussavi[1]* (farshid1@stanford.edu), Fernando Amat,[1] Luis R. Comolli,[2] Kenneth H. Downing,[2] Mark Horowit,[2] and **Harley McAdams**[3] (hmcadams@stanford.edu)

[1]Electrical Engineering, Stanford University, Stanford, Calif.; [2]Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [3]Dept. of Developmental Biology, Stanford University, Stanford, Calif.

Project Goals: The objective of this subproject is to develop algorithms to improve and accelerate analysis of electron microscopy tomography images. In particular, we make use of artificial intelligence and statistical image processing algorithms to detect and enhance biological structures within whole bacterial cell images that have high noise levels. This is a component of the overall project thrust to identify and characterize multiprotein complexes in living cells.

The advent of high throughput tomography is key to structural studies of cellular and subcellular assemblies, and remains an elusive goal. In recent years, there have been tremendous advances in the automatic acquisition of electron microscopy data[1] and generation of tomograms.[2,3] The last remaining bottleneck in the pipeline is the automatic segmentation of cellular structures. Due to severe signal to noise limitations and missing data, established image processing computer vision programs have significant challenges in these datasets. For example, apparent boundaries can be inconsistent, broken, and sometimes just nonexistent due to the missing data wedge. The segmentation of such datasets often involves days of manual effort of an expert.

To address these problems, we have developed a probabilistic framework based on dynamic Bayesian networks (DBN's) that makes significant use of context and shape as well as physical features. A weak shape prior is assumed, and gradually refined as the inference progresses through the 3D volume. So far we have obtained encouraging results on several cryo EM datasets of *Caulobacter crescentus*, achieving automatic segmentation of a membrane in less than 2 hours on a desktop computer. We are extending the model to be more robust to shape changes and different datasets.

One of our first applications of this segmentation is the study of the S-layer structure in its native state. While bacterial S-layers have been studied for over 30 years, most of the studies have been performed on isolated S-layer sheets or proteins.[4,5] We present one of the first structural studies

of S-layer in native state and show some differences with previous results.

We used datasets from cryo-electron tomography (ET) on whole cells of the gram-negative bacterium *Caulobacter crescentus* to obtain quantitative information on the S-layer structure and its interactions with the outer membrane in its native state. We employ pattern recognition and statistical analysis techniques to process efficiently the large volume of low SNR data that results from this method. Using the automatic membrane segmentation described above, we search for the S-layer in a thin volume around the model's estimated surface, by locally maximizing the characteristic S-layer hexagonal signature in the Fourier domain. Due to inevitable noise in the cryo-EM tomograms, we need to model the S-layer as a Markov Random Field to introduce smoothness constraints between neighboring subunits in the surface to minimize false local maxima.

We have identified over 4000 S-layer subunits per tomogram. Visualization reveals that the S-layer structure on the cellular surface of *C. crescentus* CB15 is not as ordered as has been observed in isolated S-layer sheets[5]. Irregularities are present on the surface which break the global[6] symmetry expected in a crystalline structure. We also observe variations on the periodicity of the lattice. We are repeating this analysis on additional *Caulobacter* strains known to have altered S-layer properties.

References

1. Zheng, S.Q. et alt. UCSF tomography: An integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction. *J. Struct. Biol.* 157:138-47, 2007.
2. Lawrence A , Bouwer, J.C., Perkins, G., and Ellisman, M.H.. Transform-based backprojection for volume reconstruction of large format electron microscope tilt series. *J. Struct. Biol.* 154:144-67, 2006.
3. F. Amat, F. Moussavi, L.R. Comolli, G. Elidan, K.H. Downing, M. Horowitz. Markov random field based automatic image alignment for electron tomography. *J. Struct. Biol.* 161:260-275, 2008.
4. Engelhardt, H. Are S-layers exoskeletons? The basic function of protein surface layers revisited. *J. Struct. Biol.* 160:115:124, 2007.
5. Smit, J. et alt. The S-layer of Caulobacter Crescentus: Three-Dimensional Image Reconstruction and Structure Analysis by Electron Microscopy. *J. Bacteriol.,* 174 (20):6527:6538 , 1992.
6. Bartesaghi, A. et alt. Classification and 3D averaging with missing wedge correction in biological electron tomography. *J. Struct. Biol.* 162:436:450, 2008.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis" (PI, Harley McAdams, Stanford University).

Control Mechanisms for Chromosome Orientation and Dynamics in *Caulobacter crescentus*

Esteban Toro,[1] Grant R Bowman,[1] Luis R Comolli,[2] Sun-Hae Hong,[1] Jian Zhu,[2] Michael Eckart,[1] Marcelle Koenig,[1] Kenneth H. Downing,[2] W.E. Moerner,[1] Thomas Earnest,[2] **Harley McAdams**,[1] and Lucy Shapiro[1]* (shapiro@stanford.edu)

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

Project Goals: We seek to develop a global and integrated view of dynamically changing structures associated with cell cycle progression and polar development. Here, we identify and characterize multiprotein complexes associated with chromosome replication and segregation, and elucidate the mechanism by which they carry out large-scale organization within the bacterial cell.

The application of state-of-the-art cell biology to the bacterial cell has yielded a paradigm shift in our understanding of these tiny cells. Our work has clearly shown that the bacterial cell is highly organized and all cell functions are integrated and regulated within the confines of a three dimensional grid. An ongoing and unanswered question is how the bacterial cell segregates its chromosomes. We have shown that in *Caulobacter*, DNA replication and segregation occur simultaneously and that the order of genetic loci on the chromosome reflects the location of these loci in the cell. Prior to the initiation of replication of the single circular chromosome, the origin region of the chromosome, which is 8 kb from the *parS* site, resides at one cell pole and the terminus at the other. Upon replication initiation, a duplicated origin region moves rapidly to the opposite pole. Thus, chromosome organization and segregation in bacteria is precise and regulated. We have identified a novel proline-rich protein, PopZ, as the anchor that tethers separated chromosome origins to the pole. PopZ localizes to the cell pole by a diffusion/capture mechanism and assembles into a filamentous network of high molecular weight oligomers. Further, PopZ interacts directly with ParB, which in turn binds *parS* sequences near to the origin of replication. Thus, at G1, the chromosome is fixed at the cell pole by the PopZ/ParB/*parS* complex. Subsequently, this interaction is released when one copy of the origin region is translocated to the opposite pole. We used chromosomal inversions and *in vivo* time-lapse imaging to show that *parS* is the site of force exertion during segregation, independent of its position in the chromosome. When *parS* is moved farther from the origin, the cell waits for it to be replicated before segregation can begin, and *parS* still forms the leading edge of the translocating chromosome. Also, a mutation in the ATPase domain of ParA halts segregation without affecting replication initiation, suggesting that ParA is the motor that drives ParB/*parS* movement. As the ParB/*parS* complex travels across the cell, PopZ accumulates at the new pole and, upon arrival, tethers

it in place. Chromosome segregation in *Caulobacter* cannot occur unless a dedicated *parS* guiding mechanism initiates movement. Together, the results show that ParB is part of a complex interaction network that controls multiple aspects of chromosome organization, including the translocation of the newly replicated chromosome and its attachment to the cell pole.

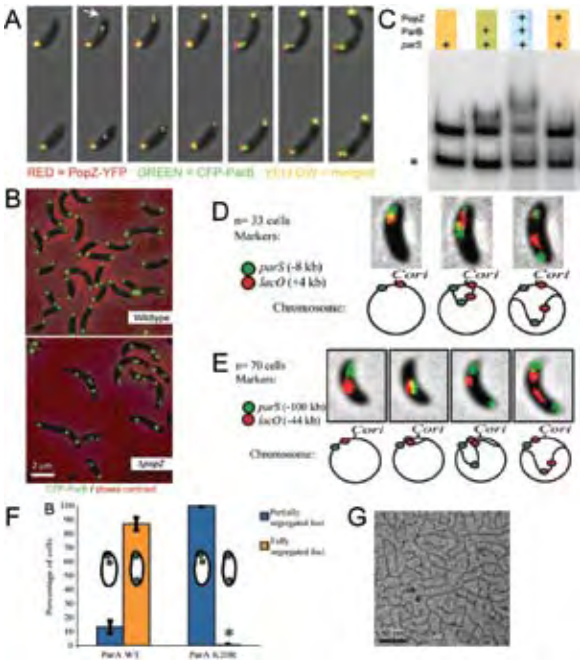


Figure. *ParB* is part of a complex network that mediates chromosome translocation and docking to the cell pole. (A) After translocation, ParB (green) co-localizes with PopZ (red) at the cell pole. Two cells were imaged by time-lapse fluorescence microscopy. (B) ParB foci (green) are no longer localized to the cell poles in a ΔpopZ strain. (C) PopZ interacts with ParB in complex with the *parS* DNA target sequence. A gel shift assay using purified components. The asterisk marks control DNA with no *parS* site. (D-E) Segregation pattern of chromosomal loci in a normal strain (D) and where the order of loci was modified by inversion (E). *Cori* indicates the origin of replication, *parS* indicates the specific ParB binding site, and *lacO* is a differentially marked locus. In both cases, ParB/*parS* is the first region to be translocated, regardless of its placement relative to *Cori* or the timing of its replication. (F) A ParA mutant disrupts chromosome segregation. After synchronization, cells were left to grow normally (WT) or stimulated to express mutant ParA K20R. A cartoon representation of the results is shown with the quantified data (bars). (G) PopZ assembles into a filamentous structure *in vitro*. Purified PopZ protein was placed on a carbon grid and visualized by transmission electron microscopy.

Optimizing the Detection of Specific Protein Complexes in *Caulobacter crescentus* at the Electron Microscopic Level

Guido M. Gaietta[1]* (ggaietta@ncmir.ucsd.edu), Ying Jones,[1] Julie Lee,[1] Masako Terada,[1] Grant Bowman,[2] Monica Schwartz,[2] Lucy Shapiro,[2] and **Mark H. Ellisman**[1] (mellisman@ucsd.edu)

[1]National Center for Microscopy and Imaging Research, Dept. of Neurosciences, University of California, San Diego, Calif.; and [2]Dept. of Developmental Biology, Stanford University, School of Medicine, Beckman Center, Stanford, Calif.

Project Goals: The work conducted by the CRBS/NCMIR team under the umbrella of the DOE-GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis" aims at generating tools for correlated light and electron microscopic analysis of microbial cells. The current efforts are focusing on (1) improving the preservation of ultrastructure and antigenicity by combining high-pressure freezing, freeze substitution, epoxy fixation and post-embedding labeling; (2) generating tools for automated tracking of macromolecular complexes in tomographic reconstructions of microbial cells.

The detection of protein complexes at the supramolecular level in microbial cells is often hindered by high sensitivity to fixatives and detergents, low preservation of the specimen ultrastructure, limited accessibility of the target protein and reduced antigenicity. To overcome these obstacles, we have developed a method that combines rapid aldehyde fixation, high-pressure freezing, freeze substitution, epoxy fixation (HPF/FS/EF) and post-embedding immunolabeling. Rapid primary aldehyde fixation followed by HPF/FS/EF avoids the bulk specimen shrinkage observed in conventionally prepared samples due to solvent dehydration and provides exceptionally well-preserved ultrastructure and morphology. The specimens embedded in the Araldite/Epon resin (we use a mixture of Araldite and Epon resins during the freeze substitution step and in the final embedding of the sample) are ultrathin sectioned (80-100 nm in thickness) and processed for immunostaining. For antigen retrieval purposes, prior to labeling the ultrathin sections are etched with a freshly made solution of saturated sodium hydroxide in 100% ethanol. The etched sections are then labeled with primary antibodies, gold-conjugated secondary antibodies, stained with uranyl acetate, Sato Lead for contrast purposes and viewed at the microscope.

This optimized labeling method was applied to evaluate the distribution of specific protein complexes related to chromosome organization and replication. Here we targeted lac arrays that were genetically inserted at different locations from the bacterium origin of replication or polar proteins

either natively expressed or in recombinant form fused to a fluorescent reporter and/or an epitope tag (HA).

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multi-institutional grant is Harley McAdams, Stanford University.

GTL

Generating, Implementing and Refining of a Custom Algorithm to Localize Ribosomes in Tomographic Reconstructions of *Caulobacter crescentus*

Rick Giuly,[1] James Obayashi,[1] Ying Jones,[1] Guido M. Gaietta[1]* (ggaietta@ncmir.ucsd.edu), Lucy Shapiro,[2] and **Mark H. Ellisman**[1] (mellisman@ucsd.edu)

[1]National Center for Microscopy and Imaging Research, Dept. of Neurosciences, University of California, San Diego, Calif.; and [2]Dept. of Developmental Biology, Stanford University, School of Medicine, Beckman Center, Stanford, Calif.

Project Goals: The work conducted by the CRBS/NCMIR team under the umbrella of the DOE-GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis" aims at generating tools for correlated light and electron microscopic analysis of microbial cells. The current efforts are focusing on (1) improving the preservation of ultrastructure and antigenicity by combining high-pressure freezing, freeze substitution, epoxy fixation and post-embedding labeling; (2) generating tools for automated tracking of macromolecular complexes in tomographic reconstructions of microbial cells.

A custom algorithm was generated and used to automatically locate centers of ribosomes in 3D volumetric images collected from a wild type strain of *Caulobacter crescentus*. The automated detection of ribosomes is challenging because (1) the 3D image of each bacterium is large (approximately one gigabyte) and therefore computationally demanding; (2) the identification of ribosomes in the image is not trivial, as they have varying intensity and shape and tend to overlap. The algorithm that we have generated consists of two steps: a **seeding step** and a **dynamic simulation step**. In the **seeding step**, the image is thresholded so that clusters of touching ribosomes tend to produce connected component regions (blobs) of a single gray scale value. For each blob, N particles with radius R are created (at the center of mass of the blob). N is proportional to the volume of the blob so that larger blobs will be seeded with more particles. In this way a group of particle "seeds" is placed at each blob. In the **dynamic simulation step**, the particles are moved so that they tend to settle into darker regions of the blob. The rule for movement of the particles is inspired by physical interaction between molecules, though highly simplified. When particles are sufficiently close they produce a repulsive force, at medium distance they produce an attractive force, and at large distances no force at all. For computational efficiency, particles within a group only produce force on other particles in their group, not on other groups associated with other blobs. In addition to force that particles exert on each other, there is a force based on the local gradient of the grayscale value in the 3D image. A "gray-scale gradient" force proportional to the magnitude of the gradient is applied to the particle so that it tends to move toward darker (minimum valued) regions. Also, at each time step the particle is displaced randomly by a small distance to help in avoiding the probability of the particle settling to local minimums.

The current iteration of our ribosome tracking software suite (source code available at cytoseg.googlecode.com) finds approximately 40% of ribosomes in the *Caulobacter c.* sample, when compared against a human tracked 3D mapping of the ribosomes. Some of the identified factors that contribute to the probability of proper tracking are resolution of the tomogram, as well as size, shape and contrast of ribosomes.

Although the algorithm does not locate all ribosomes, it gives a fully automated method of roughly visualizing the distribution. Also, when combined with manual tracing, the automated method serves to reduce the total time needed for ribosome detection. The program can run and return an output of the XYZ coordinates within hours. Comparatively, the collection of similar numbers of ribosomes by manual tracking requires several days, depending on the experience of the tracker. To improve automatic ribosome detection accuracy, in future work, we plan to develop more sophisticated approaches to segmentation incorporating prior knowledge.

This work was funded by the DOE GTL project titled "Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis". The PI of this multiinstitutional grant is Harley McAdams, Stanford University.

AMD Community Systems Understanding

GTL

Whole Community Proteomic Approaches to Decipher Protein Information from Natural Microbial Communities

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

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

The advent of integrated computational and experimental approaches for microbial genomic interrogation has afforded a unique opportunity to study microbial consortia at the molecular level. This permits a detailed examination of how environmental communities cooperate and compete for resources in natural ecosystems. To this end, we have focused on the development of experimental (LC-MS/MS) and computational approaches to characterize microbial communities in their natural settings for elucidation of how these consortia adapt and respond to their environmental pressures. Because the complexity of many natural microbial communities exceeds the current measurement capabilities of analytical techniques, it is advantageous to develop and demonstrate methodologies with a low complexity environmental microbial consortium. In this respect, the acid mine drainage (AMD) microbial system is ideal. Sufficient biomass is readily accessible to enable molecular level evaluation by a variety of genomic, proteomic, and biochemical techniques. This permits coordination of different analytical measurements on the same samples, thereby providing the ability to integrate the datasets for extraction of biological information. Furthermore, the system is moderately rich in bacterial, archaeal, and viral membership.

Whole community genomics serves as the underlying core for almost all of the subsequent measurements and evaluations of microbial consortia. The depth and quality of genome annotation, including information about bacteria (with strain variation details), archaea, and viruses, is critical for the ensuing proteomic and biochemical measurements. The recent availability of a greatly expanded genome annotation of bacteria and archaeal species in the AMD samples has provided a much richer database from which to mine proteome data. This has permitted not only a deeper level of proteome identification in the newer datasets, but also is being used to re-search older existing datasets to mine additional microbial information, especially from the archaeal members.

Examination of extensive proteome datasets from the AMD system to date has prompted the need for new experimental approaches to increase the dynamic range of proteome measurement, thereby providing a deeper and wider view of the range of protein components important for microbial community structure and function. To this end, we have explored two experimental approaches. The first approach relies on multiple mass-range scanning in the mass spectrometer to enable more extensive measurement metrics on the chromatographic time-scales. While this requires a bit more sample, the resulting datasets provide more detailed fragmentation information from a greater number of peptides in the complex mixtures, and thus provides deeper proteome information. The second approach, conducted in close conjunction with the Banfield UCB research group, relies on isoelectric focusing (IEF) fractionation of intact proteins as an additional separation step for our 2-dimensional LC/LC-MS/MS format. This has afforded not only an additional enrichment step for improving the depth of proteome coverage, but also provides an additional identification metric (i.e., pI correlation) of the proteins in each fraction. Although both of these approaches are still under evaluation in our lab, we have already observed significant improvements in proteome discovery (e.g., >6,000 proteins identified using the IEF fractionation procedure) with these and other related approaches.

In addition to improving the dynamic range of proteome measurement, we have also realized the need to better characterize the range of post-translational modifications (PTM) that impact the function of the biological ecosystem. We have focused our efforts into two specific areas. The first research area is a detailed evaluation of the extracellular proteome of the AMD system, in an effort to more fully characterize the proteins that experience the extreme environmental conditions and likely play critical roles in respiration and other vital biological activities. In particular, we have developed an integrated experimental/computational approach to investigate the range of extracellular proteins that have been processed by signal peptide cleavages. We identified ~500 signal peptide cleaved proteins with high confidence, including numerous proteins of unknown function which were further characterized by Pfam analysis. Spatial profiling of signal peptide protein expression exhibited a notable subset of proteins either conserved or divergent in their expression as a function of growth state maturity or spatial location within the mine. In particular, some cytochromes (including Cyt579) and outer membrane proteins were highly conserved across all growth states, while other putative cytochrome-related proteins were found to be highly variable across the growth states. In conjunction with the Thelen LLNL research group, we have initiated a detailed intact protein characterization of Cytochrome 579 in an effort to understand how post-translational modification and sequence variation of this protein might provide environmental tuning as a function of growth state or spatial location within the mine. The second research area involves evaluation of new computational approaches to extract PTM information from the already existing datasets. In particular, we have evaluated a new computational approach based on sequence tagging to identify a defined set of PTMs

(methylation, acetylation, and oxidation). By examining six different AMD proteome datasets, we found: a) about 20% of the peptides are modified in each sample, b) oxidative damage is much higher in the AMD system than in other bacterial isolates, and is enhanced in early growth stages, and c) flagellin and chemotaxis proteins are highly modified in late growth state biofilms, especially for L*epto III*. This suggests higher motility in late developmental growth stages.

By expanding the range of samples examined for the AMD system, as well as developing and testing new methods of increasing the dynamic range and modification identities of the proteins interrogated, we have been able to uncover a much deeper level of molecular level information about the structure and functions of natural microbial consortia.

This research sponsored by the U.S. DOE-BER, Genomics:GTL Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

GTL

Insights into the Ecology and Evolution of a Natural Microbial Ecosystem from Acid Mine Drainage using Community Genomics and Proteomics

R.S. Mueller[1]* (rmueller@berkeley.edu), V.J. Denef,[1] S.L. Simmons,[1] D. Goltsman,[1] P. Wilmes,[1] C. Belnap,[1] A.P. Yelton,[1] G.J. Dick,[1] B.J. Baker,[1] L. Comolli,[3] B.C. Thomas,[1] B. Dill,[2] L. Hauser,[2] M. Land,[2] S.W. Singer,[3] L. Kalnejais,[1] M. Shah,[2] N.C. VerBerkmoes,[2] M.P. Thelen,[3] R.L. Hettich,[2] and **J.F. Banfield**[1]

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

Project Goals: Our goal is to develop community genomic and proteomic methods to study evolutionary processes and ecological patterns of natural microbial systems. This is achieved primarily through analysis of model communities comprised of closely and more distantly related co-evolving organisms growing in an acid mine drainage environment. Community genomics methods that address both sequence variation and binning have allowed for an in-depth examination of population genetic heterogeneity, providing insights into the interplay between mutation, recombination, and transfer of mobile genetic elements and selection. Analyses of many whole community proteogenomic datasets for natural and bioreactor biofilms have enabled identification of key factors that impact organismal function, ecosystem partitioning, and dictate overall ecosystem properties.

Rigorous mining of the wealth of information produced through "omics" methods promises deeper insights into biological systems. While community genomics provides snapshots of variation from which evolutionary processes can be inferred and generates hypotheses relating to the function of ecosystems, community proteomics captures metabolic responses *in situ*, enabling hypothesis testing.

Acid mine drainage (AMD)-generating systems are characterized by hot (often >40 C), pH < 2 solutions, enriched in toxic heavy metals [mM] that are released by dissolution of sulfide minerals exposed by mining. The organisms that proliferate in AMD are well suited to cope with these extreme stresses and accelerate sulfide mineral dissolution and acid production as a byproduct of metabolism. The constrained biological diversity of this environment, and the intimate relationship between biology and chemistry, makes these ideal model ecosystems. Using the AMD system, our labs seek to develop combined cultivation-independent genomics and proteomics methods for study of microbial communities.

We reconstructed near complete genomes for the principal bacterial and archaeal members of two biofilms collected from the air-AMD interface within the Richmond mine (Iron Mountain, CA). These include typically dominant *Leptospirillum* Group II (LII) bacteria, the lower abundance and perhaps keystone member, *Leptospirillum* Group III (LIII), and Archaeal members from the *Thermoplasmatales* group and ARMAN group. Comparisons of sequences from individuals that comprise natural populations and comparative genomic analyses involving composite sequences from more distantly related but co-evolving species enable novel insights into evolutionary processes.

The fine-scale heterogeneity within LII populations was examined genome-wide to investigate intra-species evolutionary processes. The LII population is dominated by one sequence type, yet we detected evidence for relatively abundant recombining variants (>99.5% sequence identity) at multiple loci, and a few rare variants. Population genetic analyses of single nucleotide polymorphisms indicate variation between closely related strains is not maintained by positive selection, suggesting that these regions do not represent adaptive differences between strains. Thus, the most likely explanation for the observed patterns of polymorphism is divergence of ancestral strains due to geographic isolation, followed by mixing and subsequent recombination. [Simmons et al. 2008]

In order to infer possible niche specialization between community members from the same genus, we compared the genomes of LII and LIII. Both have genes for community-essential functions, including carbon fixation and synthesis of biofilm structural components, but LII is better equipped to deal with osmotic challenges. In contrast, LIII has the genes to fix nitrogen, has more chemosensing pathways, and expresses more motility proteins than LII, consistent with their observed localization within the biofilm interior, where steep chemical gradients are likely.

Further evidence for niche specialization between organisms was obtained through the examination of 28 whole community proteomes obtained from the AMD system over a four-year period. We previously reported the genomes of two LII species, the UBA and 5way types,

whose orthologous proteins differ by 5% sequence identity. Comparative genomic and proteomic analysis revealed both the type-specific expression of ortholgous genes as well as the expression of a limited subset of genes unique to each type. The most significant differentiating signal originated from genes involved in cobalamin biosynthesis and cobalamin- and S-adenosyl methionine-dependent anabolic methylation reactions. These, as well as glycine cleavage complex proteins, which generate a C_1 and ammonium, were more abundant in the 5way type, indicating that it might obtain a competitive edge by scavenging carbon sources in late developmental stage biofilms where it is most abundant. [Denef et al. 2008]

Variations in protein abundances may reflect changes in organismal activity in response to fluctuations in biotic or geochemical conditions. We used statistical methods adapted for proteomics data to determine how the community proteome expression patterns correlate with environmental factors. It was found that low abundance members of the microbial population show a strong response to fluctuations of abiotic factors, whereas the overall expression of the dominant member, LII, is less responsive. For example, the protein abundances of many *Thermoplasmatales* members and temperature correlate positively, but the activities of individual members are also affected by additional factors (i.e. *G-plasma* with [SO_4^{2-}], *A-plasma* with pH and [Ca]). Using this proteomics approach, we are able to propose niches of community members and their fine-scale relationships with specific environmental determinants, which would not be possible using comparative genomics alone. Another observation from this work is that that changes in the community composition strongly correlate with overall LII protein expression. For example, the progression from homogeneous, LII-dominated biofilms to high developmental stage films containing many other organisms coincides with a shift in the LII proteome from one biased toward ribosomal proteins and those involved in stress responses to one biased toward enzymes that produce nucleotides, amino acids and carbohydrates. Thus, the global metabolism of the most abundant organism of these biofilms is dependent on inter-species exchanges and competition, and highly resistant to changes in the prevailing environmental conditions.

These results, coupled with detailed microscopic analysis of intact biofilms from the mine showing layered microbial assemblages [Wilmes et al. 2008], inform an overarching model of ecosystem function and evolution within these microbial communities. As LII is the initial colonist of this community, due to its ability to grow alone in the extreme environment and fix carbon, establishes a stable environment for the propagation of lower abundance organisms. These organisms may perform overlapping roles with the ecosystem, but appear to be adapted to well-defined environmental conditions. Functional capacity and adaptations to differing conditions are driven by genetic heterogeneity and exchange between close relatives, in addition to changes in the regulation of key metabolic pathways in response to the external environment.

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

GTL

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

Christine Sun[1]* (christine_sun@berkeley.edu), Christopher Belnap,[1] Brian Thomas,[1] Kimberly Pause,[2] Paul Wilmes,[1] Daniela Goltsman,[1] Chongle Pan,[3] Anders Andersson,[1] Robert Hettich,[3] Nathan VerBerkmoes,[3] Mya Breitbart,[2] and **Jill Banfield**[1]

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

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

Despite their tiny size, viruses represent the most abundant entities on Earth. Bacteriophages and archaeal viruses (here collectively referred to as viruses) have the ability to shape the composition and functionality of microbial communities through predation of their hosts and lateral gene transfer. Since viruses can play central roles in microbial evolution, it is important to understand the dynamic interactions between viruses and their microbial hosts. Yet, few studies have examined virus-host dynamics and the diversity of viruses in natural populations. While various metagenomic studies have allowed glimpses into functional gene diversity in viral populations, no cultivation-independent method has existed to link a particular virus with its hosts prior to the identification of CRISPRs (clustered regularly interspaced short palindromic repeats) as adaptive, rapidly evolving microbial immune systems. Simultaneous genomic analysis of the CRISPR loci of natural bacterial and archaeal populations and the virus populations that they target *provides a new route to study virus–host interaction dynamics* [Tyson *et al.* 2007; Andersson *et al.* 2008]. The relatively low diversity microbial community in acid mine drainage (AMD) from Iron Mountain, CA, allows the opportunity to comprehensively examine the dynamics between CRISPR regions and viruses. The findings from this system are likely generally applicable and directly relevant to the DOE mission through the implications for the maintenance of stable

* Presenting author

biotechnologies, including bioremediation and industrial bioenergy production.

A CRISPR locus consists of repeats separated by short unique spacer sequences that originate from viral genomes and other mobile elements. The *cas* (CRISPR-associated) genes, which are found adjacent to CRISPR loci, encode proteins that are key to the functioning of the immune system. Microbial immunity to viruses derives from having a CRISPR spacer sequence exactly matching a viral sequence [see Barrangou *et al.* 2007]. Since spacers incorporate into the CRISPR locus in one direction, the CRISPR region provides a record of recent viral exposure. We have shown that CRISPR loci within microbial genomes assembled from AMD community genomic datasets are extremely dynamic genomic regions that display evidence of extensive loss and gain of spacers. Using 454 FLX sequencing of CRISPR loci from two *Leptospirillum* group II bacterial populations, we recovered 521,203 total spacer sequences, of which 16,144 were unique. Rarefaction curves built from these CRISPR spacers show no approach to saturation. The diversity in spacer sequences within host populations is consistent with evolution of the spacer complement on microbial generational time scales.

The viral-derived spacer sequences have allowed us to identify previously unassigned sequences in AMD community genomic datasets as viral and to assemble the viral sequences into 5 near-complete genomes (AMDV1-AMDV5), enabling unprecedented analysis of heterogeneity within natural virus populations [Andersson *et al.* 2008 and unpublished]. Recombination is apparent within all viral populations and in one population (AMDV2) it is sufficiently extensive to serve as a means of avoiding the short sequences of CRISPR spacers.

Quantitative proteomic analyses of replicate, laboratory-grown communities suggest a stochastic nature of viral infection within the AMD biofilms. Three replicate community biofilm samples each contained different viral types (AMDV1 vs. AMDV3 vs. AMDV4), with few to no proteins identified from the other AMD viral types. Given that communities were grown from the same starting inoculum, we infer that viral blooms can be unpredictable and spatially heterogeneous. This may be a consequence of rapid evolution within the "clouds" of host resistance and viral sequence diversity. There is also evidence for spatial and temporal heterogeneity in the host response for *Leptospirillum* group II in these laboratory-grown biofilms. Five of the Cas proteins from *Leptospirillum* group II were ~60 times more abundant in biofilm-associated cells than planktonic cells, yet no such distinction was observed in a replicate community grown within a separate laboratory reactor. This, and related findings for natural communities, indicate strong regulation of the CRISPR/*cas* system. Heterogeneity in viral populations is also apparent at larger scales within the field site. For example, distinct AMDV1 sequences are only found in certain community genomic datasets and identification of Cas proteins from host populations is quite variable.

Aligning CRISPR spacer sequences to the viral genomes reveals information about the still poorly understood CRISPR mechanism. For example, 1,235 unique spacers derived from 454 sequencing matched AMDV1, the bacteriophage with *Leptospirillum* group II and III as hosts. Different spacers target different variants of the same viral type. In some cases, these spacers overlap with each other and even target different variants in the same genomic region. Also, plasmids have been found to contain CRISPR loci. Plasmid populations reconstructed from AMD community genomic datasets have CRISPRs with spacers that target the *cas* sequences of other plasmid CRISPRs, presumably to silence that plasmid's immune system. This illustrates a previously unknown phenomenon: CRISPRs may be involved in plasmid versus plasmid competition.

All evidence from the AMD system indicates that both CRISPRs and viruses are co-evolving rapidly, in an "arms-race" that requires continual acquisition of new spacers to target viral sequences that have been modified by mutation and sequence shuffling. The features of this emerging view of virus-host interactions are consistent with findings published from parallel pure culture laboratory studies conducted by other groups [Barrangou, Horvath and others]. We infer that the dynamic balance in infectivity and resistance is a general feature of virus-host interactions, a result that underlines the importance of viruses in shaping the structure and functioning of microorganisms in ecosystems of both natural and biotechnological importance.

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

GTL

Shotgun Proteomics and *De Novo* Sequencing for the Detection of Viral Signatures in Natural Microbial Communities

Nathan VerBerkmoes[1]* (nve@ornl.gov), Chongle Pan,[1] Brian Dill,[1] Patricia Carey,[1] Jacque Young,[1] Doug Hyatt,[1] Manesh Shah,[1] Kim Pause,[2] Christine Sun,[3] Brian Thomas,[3] Paul Wilmes,[3] Robert L. Hettich,[1] Mya Breitbart,[2] and **Jillian Banfield**[3] (jbanfield@berkeley.edu)

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

Project Goals: The aim of this project is to develop integrated genomic and proteomic (proteogenomic) methods to study virus-host dynamics in bioreactor-grown and natural microbial biofilm communities. We will utilize simultaneous genomic analysis of CRISPRs (the adaptive microbial immune system) and viruses to examine the co-evolution and dynamics of viruses and their microbial hosts. Due to virus-host specificity, viruses can change the microbial community structure by enriching for resistant strains. Proteomics will be used to examine metabolic interplay via monitoring of protein production before,

during, and after viral infection, while genomic analysis of CRISPRs will be used to correlate these findings with changes in community structure. Because viral predation can undermine microbially-based technologies such as microbial fuel cells, bioethanol production, and environmental remediation, methods to study microbe-viral interactions are widely relevant to DOE missions.

Microorganisms comprise the majority of extant life forms and play key roles in a wide variety of health and environmental processes, yet little is known about the nature and driving forces of their diversification. Although the roles of viruses in microbial evolution are widely recognized, neither the details of viral-microbial interactions nor the impact of virus on microbial community structure are well understood. Community genomic and proteomic (proteogenomic) methods have been established for analyzing the roles and activities of uncultivated bacteria and archaea in natural multi-species consortia (Ram, *Science* 2005, Lo *Nature* 2007, Wilmes, *ISME*, 2008, VerBerkmoes, *ISME*, 2008). Notably lacking are methods for monitoring viral activity in communities, tracking virus predation, and determining the consequences of viral predation for ecosystem structure. Our work focuses on acid mine drainage (AMD) biofilms because of their relative simplicity and established utility as a model system for development of methods for cultivation-independent analyses. Our goal is to develop proteomics and informatic techniques to characterize the microbial response to viral attacks as well as to identify the viral signatures in natural microbial communities.

One of our first goals was examine the distribution and spectral evidence of viral proteins and peptides across all AMD biofilm proteomes analyzed to date (this includes over 8 full proteomes (multiple fraction per proteome) and 40 sub-proteomes (only whole cell fractions). All proteome were re-analyzed with a new database that contained the consensus predicted protein sequences for four different contigs hypothesized to contain phage proteins. *The central hypothesis to this part of the study is there will be conserved viral proteins and peptides from those proteins that can be found in multiple sampling sites.* Furthermore the MS/MS spectra, from these conserved viral peptides, should act as molecular fingerprints for viruses in the AMD system. To support the hypothesis it will be necessary to find unique peptides, with high quality spectra, that were reproducible over numerous locations in the mine. The first step of this study was to compile all viral proteins into a central dataset. Using the spectral counts from the database results, those proteins annotated as 'viral' were compiled into a database which includes the annotated name, the location(s) in which they were present, the number of the most abundant spectra and the fraction from which the highest spectral count was found. We were able to find a number of conserved viral peptides across most of the biofilms and manually validated MS/MS spectral evidence for the expression of viral proteins in all biofilms analyzed to date. One of the major questions though is how many viral peptides and proteins are we missing due to strain variation within the viral genomes? One possible solution to this problem is the use of strain resolved protein databases instead of consensus protein databases. Our current approach was to use a six frame translation of all the viral proteins including all of their possible variants. Currently, we are examining the results from this analysis.

The possibility exist that no matter how many different viral strain resolved proteins, that are in a database, that the virus proteins will evolve differently in different samples; thus making it impossible to identify the viral proteins with the typical database searching algorithms. Alternative advanced analytical and informatic methodologies are needed to identify rapidly evolving viral proteins in complex microbial communities. *De novo* sequencing programs aim to derive complete or partial amino acid sequences from MS/MS scans without complete information from protein sequence databases. Although *de novo* sequencing approaches seem straightforward, they have not been widely applied in proteomics due to low data quality and software limitations. Thus one of our major efforts is to improve *de novo* sequencing methods to specifically be able to identify viral peptides, even if they have changed 1-2 amino acids for every ten amino acids of protein sequence. We have three major subtasks to accomplish this goal: 1) Develop and test improved analytical methods to create better mass spectral data for the algorithms to use; 2) Test all available commercial and public *de novo* and sequencing tagging algorithms; and 3) Develop a new *de novo* sequencing algorithm that can directly use the improved MS data.

The development of better mass spectral data is critical for de novo sequencing in our view. We believe that the biggest hindrances to high quality de-novo sequencing are low quality MS data. Our approach is to create datasets using our normal 2d-LC-MS/MS methods that contain three specific analytical data for each peptide sequenced. Our goal for every peptide is to acquire 1) a high resolution parent peptide mass spectra with mass accuracies of 1-5ppm with internal calibrations; 2) a high resolution CID (collisional induced dissociation) spectra 3) a high resolution ETD (electron transfer dissociation) spectra. We have currently established steps 1 and 2 on LTQ-Orbitraps in automated 24 hour runs, we have established linked CID and ETD spectra of the same peptide, at low resolution on LTQs and we feel that as soon as the LTQ-Orbitraps are upgraded with ETD; this should be straightforward to obtain all three data points for all peptides analyzed. The major downside of this approach is duty cycle; in general shotgun proteomics every peptide can be analyzed in 100-200 milliseconds. This integrated approach of linked high resolution full scans, CID, and ETD scans will take at least 1-2 second per peptide. One solution to this problem that we have been exploring is the use of 2d-LC-MS/MS with multiple mass range scanning.

We have compared a number of *de novo* programs. The datasets for comparisons was a standard microbial isolate proteome (*R. palustris*) and an AMD proteome. To mimic unknown strain variation we eliminated the genome of the dominate strain and used a related strain as the template. We evaluated the following: speed, accuracy (total number of amino acids in a string it could find), scoring method (how accurate is the scoring method, do high scores always equal correct identifications, ability to use a diversity of mass

spectral datasets, dynamic range of identifications and general ease of use. After testing we conclude that two of the newer algorithms provide the best results but each had positive and negative aspects to them. PEAKS is a commercial software, it had medium run speed, the best accuracy (*it was the only program we have seen that can get 100% of the peptide correct*), was straightforward to operate and could handle data from multiple MS platforms. DirecTag (available from Vanderbilt University) is a very fast algorithm, but it only provides short tags, its scoring scheme is not as accurate as PEAKS, but it provides tags for many more spectra than PEAKS. From these initial studies it was clear that no one given algorithm, at this point, can provide all the answers or is the best for all potential experiments. Clearly the integration of algorithms and the use of higher quality mass spectral data discussed above will be necessary.

High-resolution tandem mass spectra can now readily be acquired with hybrid instruments, in high-throughput shotgun proteomics workflows, as discussed above. *We have developed a new de novo sequencing algorithm, Vonode, has been developed specifically for such high-resolution tandem mass spectra.* To fully exploit the high mass accuracy, sparse noise, and low background of these spectra, a unique scoring system is used to evaluate sequence tags based mainly on mass accuracy information of fragment ions. Consensus sequence tags were inferred for 11,422 spectra with an average peptide length of 5.5 residues from a total of 40,297 input spectra acquired in a 24-hour proteomics measurement of *R. palustris*. The accuracy of inferred consensus sequence tags was 84%. The performance of Vonode was shown to be superior to the PepNovo v2.0 algorithm, especially in term of the number of *de novo* sequenced spectra in a head to head test. Currently we are developing a fully integrated software package for this software and testing it along side PEAKS and DirecTag.

This research sponsored by the U.S. DOE-BER, Genomics:GTL Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

GTL

Characterization of Viruses from an Acid Mine Drainage System

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

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

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

Viruses are the most abundant biological entities on the planet, playing important roles in biogeochemical cycling, horizontal gene transfer, and defining the community composition of their hosts. We are only beginning to understand the identity, diversity, and ecology of viruses in the environment. Community analysis through metagenomic sequencing has proven to be a useful tool for examining viruses in a variety of environments. Studies have revealed diverse environmental viral communities with a high degree of novelty. Previous community analyses of the acid mine drainage (AMD) microbial biofilm communities at our study site (Richmond Mine at Iron Mountain, CA, USA) have demonstrated that this is a low-diversity system that can offer unique insight into virus-host interactions by concomitant genomic sampling of virus populations and clustered regularly interspaced short palindromic repeat (CRISPR) loci (See Sun *et al.* poster).

We have optimized methods to isolate the viral community from two distinct types of AMD biofilm, and used microscopy and metagenomic sequencing to characterize the viral community. A combination of mechanical disruption, filtration, and density-dependent centrifugation successfully enabled purification of viruses from the biofilm. The purity of the fractions was examined using epifluorescence microscopy and PCR amplification of bacterial and archaeal 16S genes. Several different viral morphologies were observed by electron microscopy, including icosahedral capsids typical of bacteriophages, and lemon and rod-shaped morphologies similar to archaeal viruses found in other extreme environments. To verify the presence of viruses in the sample, viral sequences from previously obtained community metagenomic datasets were used to design primers to specifically target AMDV1, a bacteriophage of *Leptospirillum* groups II and III. Sequences generated from these PCR clone libraries exhibited extensive single nucleotide polymorphisms among the AMDV1 virus population.

Metagenomic sequencing was also used to characterize the purified viral community. Preliminary screening demonstrated that approximately 80% of the sequences had similarity to known AMD viral or putative viral sequences, confirming the success of the virus purification method. This may indicate that there is overall low virus diversity at the population level, which is consistent with the low

microbial diversity identified in previous studies in this system. Additionally, our sample was collected in 2007, more than 3 years after the samples were collected from which the genomic database was developed. This suggests that there may be long-term stability of virus populations in the AMD system, making it an ideal place to study viral evolution in a natural setting. An important next step is to examine purified virus DNA from the AMD biofilm in order to compare CRISPR loci to coexisting viruses and identify viruses not targeted by the CRISPR system.

Finally, we are working with *Leptospirillum* group II isolates, originally cultured from the mine for a previous study. Efforts are underway to infect these cultures with viruses isolated from the biofilm samples, with a goal of isolating virus-host systems for laboratory manipulation. We are also working to characterize the vesicles produced by the *Leptospirillum* group II isolates, which are found in culture as well as in the natural environment. Insight gained from these experiments may help in eliminating vesicles and any associated nucleic acids from the purified viral fraction. Additionally, we will examine the role that these vesicles may play in gene transfer.

GTL

Microbially Mediated Transformation of Metal and Metal Oxide Nanoparticles

M.J. Doktycz[1]* (doktyczmj@ornl.gov), D.A. Pelletier,[1] C.K. McKeown,[1] S.D. Brown,[1] B. Gu,[2] W. Wang,[2] D.P. Allison,[1,3] D.C. Joy,[3,4] and T.J. Phelps[1]

[1]Biosciences Division and [2]Environmental Sciences Division, [3]University of Tennessee, Knoxville, Tenn.; and [4]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The goal of the project is to understand the potential fate and transport of engineered materials in the environment. To do so, the interactions between engineered nanoparticles and bacterial systems are being investigated. The technical approach involves three aims. First, the preparation of well-characterized metal and metal oxide nanoparticles is being pursued. Well-defined nanoparticle structures are essential for drawing accurate conclusions on how physical and chemical characteristics influence microbial response. The response of various microbial species to these materials will be assessed by conventional bacterial growth, viability, and gene expression assays (Aim 2) to determine exposure tolerance and molecular system activation. These efforts coordinate closely with advanced imaging studies (Aim 3) to visualize, and chemically and physically characterize, microbial reactions to nanoparticle exposure. There is close coordination between all aims in order to gain a clear understanding of how nanoparticle composition, size, and concentration affect microbial response. Correlations between microbial response and the physical and chemical properties of the nanoparticles are sought. This information will provide a foundation for understanding the potential fates of these materials in the environment, for guiding the development of effective nanoparticle-based technologies, and for understanding how microbial systems adapt to these exposures.

Engineered nanostructures have a central role in energy conservation strategies and economic growth. One of the most significant impacts of engineered nanostructures is for effecting heterogeneous catalysis as required for fuel transformation, energy storage, polymer production and chemical synthesis. Metal and metal oxide nanoparticles are often used. They possess high surface areas and the ability to selectively mediate chemical transformations. The size and composition of the particle affects performance and may similarly affect nanoparticle fate and transport in the environment. The transformation of such nanoparticle catalysts in the environment is likely to proceed through interactions with bacteria. Several bacterial species are well known to interact with nanoparticles. Nanoparticle production, nanoparticle toxicity, nanoparticle binding and incorporation with bacteria have all been observed. However, basic knowledge that would allow prediction of the probable interaction between an engineered nanoparticle and bacteria is lacking. Our efforts seek to quantify and characterize interactions between engineered metal and metal oxide nanoparticles and selected microbial species. The effect of size and chemical composition of nanoparticles that are currently considered for various applied uses are being studied. Initial efforts are focused on the effects of cerium oxide nanoparticles on the growth, viability and genetic response of *E. coli*. Well-characterized CeO_2 nanoparticles of various sizes have been prepared and presented to bacterial cells in a dose dependent manner. Advanced imaging techniques are used to evaluate the binding and fate of the nanoparticles and the bacterial cell. The results of these studies will provide a basis for understanding how nanoparticle size and composition influence their interactions with microorganisms, and how microorganisms may alter the fate and transformation of engineered nanoparticles in the environment.

Systems Biology Strategies and Technologies for Understanding Microbes and Microbial Communities

Genomic and Proteomic Strategies

GTL

Profiling Microbial Identity and Activity: Novel Applications of NanoSIMS and High Density Microarrays

Eoin Brodie[1]* (elbrodie@lbl.gov), Jennifer Pett-Ridge[2] (pettridge2@llnl.gov), Peter Weber,[2] Gary Andersen,[1] Meredith Blackwell,[3] Nhu Nguyen,[4] Katherine Goldfarb,[1] Stephanie Gross,[3] and Paul Hoeprich[2] (hoeprich2@llnl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Lawrence Livermore National Laboratory, Livermore, Calif.; [3]Louisiana State University, Baton Rouge, La.; and [4]University of California, Berkeley, Calif.

Project Goals: The identification of microorganisms responsible for specific processes remains a major challenge in microbial ecology, one that requires the integration of multiple techniques. We propose to address this goal by developing a new methodology, "Chip-SIP"; combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) analyses, and linking the identity of microbes to their functional roles. Building upon the stable isotope probing approach (SIP), we will isotopically label microbial nucleic acids by growing organisms on ^{13}C and ^{15}N enriched substrates. Extracted RNA will be hybridized to a newly engineered high-density oligonucleotide microarray with a conductive surface and higher reproducibility relative to traditional microarrays. These advances in array surface chemistry will allow us to successfully analyze arrays by NanoSIMS, generating isotopic and elemental abundance images of the surface, and thereby indicating which organisms in complex consortia utilized the isotopically labeled substrate. Our first application of this new method will assign function to complex microbial community members dwelling in the hindgut of the wood-eating passalid beetle, *Odontotaenius disjunctus*. This microbial community represents a naturally-selected highly-efficient lignocellulose degrading consortium. Understanding the microbial processes by which wood-ingesting insects derive energy may aid large-scale conversion of lignocellulosic biomass into biofuels.

Identification of microorganisms responsible for specific metabolic processes remains a major challenge in environmental microbiology, one that requires the integration of multiple techniques. The goal of this project is to address this challenge by developing a new methodology, "Chip-SIP", combining the power of re-designed oligonucleotide microarrays with nano-scale secondary ion mass spectrometry (NanoSIMS) to link the identity of microbes to their metabolic roles.

This concept involves labeling of microbial nucleic acids following incubation with a stable isotope-labeled compound (e.g. ^{13}C-cellulose or $^{15}N_2$). Extracted RNA is hybridized to a newly engineered high-density oligonucleotide microarray with a conductive surface and higher reproducibility relative to traditional glass/silane microarrays. These advances in array surface chemistry allow successful NanoSIMS analysis of the microarray surface with hybridized nucleic acids, generating isotopic and elemental abundance images of the array surface, and thereby indicating the identity of organisms incorporating the isotopically labeled substrate.

To date, a cyclo-olefin co-polymer plastic (COP) was identified that meets our requirements for these new microarrays (opacity comparable to glass, minimal autofluoresence, adequate hardness and temperature stability to enable surface coating processes). These COP slides were coated with ~400 angstroms ITO (indium tin oxide) and the surfaces functionalized with alkyl phosphonates. We synthesized highly reproducible oligonucleotide probe features on these alkyl phosphonate-ITO surfaces using our NimbleGen microarray synthesizer unit and successfully hybridized DNA and RNA. These alkyl phosphonate-ITO surfaces demonstrated 10X lower background and 10X higher signal:noise compared to traditional arrays. Furthermore these new ITO arrays (unlike silane arrays) may be repeatedly hybridized/stripped while maintaining high performance, allowing reuse and reduced costs (patent application filed). In preliminary NanoSIMS analyses of the ITO array surfaces, we have demonstrated the ability to analyze individual oligonucleotide probe features. We have improved environmental RNA extraction and biotin end-labeling procedures and calibrated hybridization protocols for ^{13}C-labeled RNA from pure cultures and mixed bacterial consortia. Incubations with our model organism, the wood-eating passalid beetle, *Odontotaenius disjunctus*, are ongoing and include feeding experiments with ^{13}C-labeled cellulose and lignin and incubation under an atmosphere of $^{15}N_2$ to identify the location and identity of lignocellulose-degrading and nitrogen-fixing bacteria respectively within this beetle hindgut.

GTL

NanoSIP: Functional Analysis of Phototrophic Microbial Mat Community Members Using High-Resolution Secondary Ion Mass Spectrometry

Luke Burow,[1,2*] Dagmar Woebken,[1,2*] Lee Prufert-Bebout,[2] Brad Bebout,[2] Tori Hoehler,[2] Jennifer Pett-Ridge,[3] Steven W. Singer,[3] Alfred M. Spormann,[1] and **Peter K. Weber**[3] (weber21@llnl.gov)

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

We are developing a new technique—nanoSIP—to measure nutrient uptake and assimilation at the single cell level. The method combines *in situ* phylogenetic and immuno-labeling methods with stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) analysis to link microbial identity to function in complex microbial communities. We are using this new method to study the hydrogen ecology of phototrophic microbial mats. Our studies seek to document the ecological factors that affect the pathways and efficiency with which solar energy is captured, stored in chemical form, and either dissipated within the mat or released as a usable resource. In this work, we demonstrate the application of this method to study key primary producers in a layered, dihydrogen (H_2)-evolving, phototrophic microbial mat from Elkhorn Slough, CA.

Life in phototrophic microbial mats is governed by the diel cycle. In current studies we are investigating the biogeochemical parameters of the Elkhorn Slough microbial mat during a diel cycle to learn more about metabolic processes occurring in this mat, which are responsible for H_2 production. Distinctly contrasting biogeochemistry can be observed within the surface H_2 producing layer over the cycle, with high fluxes of H_2 occurring in the dark anoxic period. H_2 flux is thought to be driven by oxygen-sensitive fermentation and N_2 fixation. These processes cease during the day due to photosynthesis, which generates oxic conditions in the mat. During the diel experiment, nitrogen fixation was measured via the acetylene reduction assay, oxygen profiles were obtained using microelectrodes and hydrogen concentrations were measured using a flux chamber. Additionally we measured photosystem II efficiency using pulsed amplitude modulated fluorometry. Mats were incubated with $H^{13}CO_3^-$ and $^{15}N_2$ to isotopically label newly fixed C and N (Fig. 1). We are using nanoSIP and Catalyzed Reporter Deposition- Fluorescence In Situ Hybridization (CARD-FISH) to identify the microorganisms that fix C and N and quantify their metabolic activity and thus their importance to primary production in microbial mats. Our long-term goal is to link these parameters to the investigation of diel metabolic dynamics of microorganisms within the mats by using a global metatranscriptomic strategy.

Figure 1. Light micrograph and nanoSIP images of *Lyngbya* spp. in a microbial mat showing net carbon and nitrogen uptake. *Lyngbya* is numerically abundant in the Elkhorn microbial mat and these data show that it is an active C and N fixer in this mat. By contrast, another numerically important cyanobacteria, *Microcoleus* (not shown), fixes significant levels of C but not N in this mat.

GTL

NanoSIP: Linking Microbial Phylogeny to Metabolic Activity at the Single Cell Level Using Element Labeling and NanoSIMS Detection

J. Pett-Ridge[1]* (pettridge2@llnl.gov), S.W. Singer,[1]* S. Behrens,[2] T. Lösekann,[3] W.-O. Ng,[2] Dagmar Woebken,[2] Rick Webb,[4] B. Stevenson,[5] D. Relman,[3] A.M. Spormann,[2] and **P.K. Weber**[1] (weber21@llnl.gov)

[1]Lawrence Livermore National Laboratory, Livermore, Calif.; [2]Dept. of Civil and Environmental Engineering and [3]Depts. of Microbiology & Immunology and of Medicine, Stanford University, Stanford, Calif.; [4]Centre for Microscopy and Microanalysis, University of Queensland, Brisbane, Australia; and [5]Dept. of Botany and Microbiology, University of Oklahoma, Norman, Okla.

We are developing a new technique—nanoSIP—to measure nutrient uptake and assimilation at the single cell level. The method combines in situ phylogenetic and immuno-labeling methods with stable isotope probing (SIP) and nanometer-scale secondary ion mass spectrometry (NanoSIMS) analysis to link microbial identity to function in complex microbial communities. We are using this new method to study the hydrogen ecology of phototrophic microbial mats. Our studies seek to document the ecological factors that affect the pathways and efficiency with which solar energy is captured, stored in chemical form, and either dissipated within the mat or released as a usable resource.

One of the methods we are developing uses rRNA-based fluorescence in situ hybridization combined with an elemental label (EL-FISH). This approach allows simultaneous phylogenetic and SIP imaging in the NanoSIMS. Fluorine or bromine atoms were introduced into cells via 16S rRNA-targeted probes, which enabled phylogenetic identification of individual cells by NanoSIMS elemental imaging. To overcome the natural halogen backgrounds, we used the catalyzed reporter deposition (CARD) -FISH technique

with halogen-containing fluorescently labeled tyramides as substrates for the enzymatic tyramide deposition. The relative cellular abundance of fluorine or bromine after EL-FISH exceeded natural background concentrations by up to 180-fold, and allowed us to distinguish target from non-target cells in NanoSIMS fluorine or bromine images. The method was optimized on axenic cultures and applied to a dual-species consortium (filamentous cyanobacterium and heterotrophic alpha-proteobacterium) and complex microbial aggregates from human oral biofilms. We have also conducted a multi-factorial experiment to test the effects of the EL-FISH technique on isotopic enrichment in both pure and mixed cell cultures. We found significant effects of fixation reagent and CARD-FISH that must be accounted for when interpreting isotopic enrichment data.

The other visualization method we are developing is based on immuno-labeling of functional proteins with nanoparticles. The nanoparticles can be imaged directly with the NanoSIMS to enable simultaneous imaging of functional protein and SIP imaging. The advantage of this approach over correlated TEM-NanoSIMS studies is that it eliminates the TEM step, it allows for thick-section analysis, and electron dense phases in the sample are readily distinguish by elemental composition in the NanoSIMS. We have succeeded in imaging Au nanoparticles in the NanoSIMS at standard antibody densities. We are working on developing this method in axenic *Lyngbya* cultures and complex microbial communities, focusing on nitrogen fixation in cynobacteria.

GTL

Production of Extracellular Polymeric Substances by a Natural Acidophilic Biofilm

Y. Jiao,[1]* (jiao1@llnl.gov), M. Shrenk,[2] G. Cody,[2] **J.F. Banfield**,[3] and **M.P. Thelen**[1]

[1]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.; [2]Geophysical Laboratory, Carnegie Institution of Washington, Washington, D.C.; and [3]Dept. of Environmental Science, Policy, and Management, University of California, Berkeley, Calif.

Project Goals: This is part of a project led by JF Banfield at the University of California at Berkeley. In this subproject (MP Thelen, PI), our goals are to identify and characterize the functions of biofilm constituents isolated from a low-diversity community of microbes engaged in iron oxidation. Biofilms are collected from an EPA Superfund site characterized by extremes of high heat and acid pH, leading to acid mine drainage pollution. Both proteins and extracellular polymeric substances (EPS) from biofilms at different growth stages are under current investigation, towards a comprehensive understanding of the ecology and organization of this acidophilic community.

Biofilms are ubiquitous in nature. In most cases, biofilms facilitate the formation of specialized microbial communities adhering to natural or artificial surfaces through the production of extracellular polymeric substances (EPS). The embedding of microorganisms within complex polysaccharides and other biopolymers promotes microbial assemblages, cell adhesion, and community protection from hostile environmental conditions. As major structural components, EPS provides spatial organization and structural stability to the microbial community. The exact composition of EPS varies substantially between different biofilms, and it remains a considerable challenge to provide a complete biochemical profile for interpreting how changes in EPS constituents affect community organization and development. Here we present a framework of EPS analysis whereby the correlation between EPS composition and potential microbial functions can be explored.

As part of our ongoing proteogenomic investigation of acidophilic communities collected in an acid mine drainage site (Richmond Mine, Iron Mountain, California; see Mueller et al. poster), we examined the properties of EPS as a function of biofilm development. EPS was extracted by ethanol precipitation after biofilms were homogenized in a sulfuric acid solution. The quality and quantity of EPS obtained from early growth stage (GS1) and mid-stage (GS2) biofilms were evaluated. More than twice the amount of EPS was obtained from GS2 compare to that of GS1, with approximately 15 and 40 mg of EPS per gram of dry weight for GS1 and GS2, respectively. Chemical composition analysis indicated the presence of carbohydrate and heavy metals, and minor quantities of protein and DNA, although the relative concentration of each component varied between the two EPS samples. EPS from the GS2 biofilm contains significantly higher concentrations of carbohydrate and heavy metals compared to GS1. Glycosyl composition analysis indicates that both EPS samples are composed primarily of galactose, glucose, heptose, rhamnose and mannose, and the relative amount of individual sugars varies substantially with developmental stage (see Table 1).

Table 1. Glycosyl Composition Analysis of GS1 and GS2 samples

Glycosyl residue	GS1 (Mol%)	GS2 (Mol%)
Galactose (Gal)	51.9	18.9
Glucose (Glc)	21.2	16.7
Heptose (Hep)	12.2	33.0
Rhamnose (Rha)	11.1	13.3
Mannose (Man)	8.7	8.0
Xylose (Xyl)	n.d.	2.6
3 Deoxy-2-manno-2 Octulsonic acid (KDO)	n.d.	7.8
3OH C16 Fatty Acid	n.d.	+
N Acetyl Glucosamine (GlcNAc)	n.d.	+

Table 2. Comparison of glycosyl linkage analysis of GS1 and GS2 samples.

Glycosyl Residue	GS1 (mol%)	GS2 (mol%)
terminally linked hexofuranosyl residue (t-hexf)	21.2	10.9
4-linked glucopyranosyl residue (4-Glc)	13.1	8.2
terminally linked rhamnopyranosyl residue (t-Rha)	12.2	15.5
terminally linked heptopyranosyl residue (t-Hep p)	9.9	10.6
3,4-linked galactopyranosyl residue (3,4-Gal)	8.6	3.7
terminally linked mannopyranosyl residue (t-Man)	7.2	12.3
2-linked mannopyranosyl residue (2-Man)	4.6	9.6

Additionally, carbohydrate-linkage analysis reveals multiply linked heptose, galactose, glucose, mannose and rhamnose (see Table 2), similar to the complex, branched polysaccharides found in plant cell walls, and perhaps also expected in an environmental biofilm matrix. Interestingly, much of the glucose measured is in the 4-linked form. Consistently, solid-state NMR analysis of EPS samples indicates that up to 25% of the EPS is cellulose (β1→4 glucan).

Besides providing a structural element in biofilms, EPS in acid mine drainage microbial communities may offer protection from toxic heavy metals through diffusion limitation, and facilitate nutrient flux between different members within the community by acting as a carbon source/sink. In addressing the covariance pattern of EPS composition with biofilm developmental stages and microbial processes, we hope to gain a fundamental understanding of how EPS functions in natural microbial communities.

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

GTL

Soil Community Metagenomics at the DOE's Climate Change Research Sites

Cheryl Kuske[1]* (kuske@lanl.gov) and the FACE soil metagenomics working group: Gary Xie,[1] John Dunbar,[1] Larry Ticknor,[1] Yvonne Rogers,[1] Shannon Silva,[1] La Verne Gallegos-Garcia,[1] Stephanie Eichorst,[1] Shannon Johnson,[1] Don Zak,[2] Rytas Vilgalys,[3] Chris Schadt,[4] Dave Evans,[5] Patrick Megonigal,[6] Bruce Hungate,[7] Rob Jackson,[3] David Bruce,[1,8] and Susannah Tringe[8]

[1]Los Alamos National Laboratory, Los Alamos, N.M.; [2]University of Michigan, Aspen FACE site, Ann Arbor, Mich.; [3]Duke University, Pine FACE site, Durham, N.C.; [4]Oak Ridge National Laboratory, Sweetgum FACE site, Oak Ridge, Tenn.; [5]Western Washington University, Desert FACE site, Bellingham, Wash.; [6]Smithsonian Institution, Estuary FACE site, Washington, D.C.; [7]Northern Arizona University, Scrub Oak OTC site, Flagstaff, Ariz.; and [8]DOE Joint Genome Institute, Walnut Creek, Calif.

Project Goals: Our overarching vision is to establish a science strategy and implementation pipelines to obtain a comprehensive, field-scale understanding of the structure of soil microbial communities and their functional processes that play critical roles in terrestrial carbon sequestration and response to climate change perturbations. Our three goals provide complementary science and technology information to achieve this overall vision. (1) Understand the impacts of long-term elevated CO_2 and other environmental factors (ozone, nitrogen interactions) on the structure and activities of soil microbial communities, at the DOE's six Free Air CO_2 Enrichment (FACE) and Open Top Chamber (OTC) experimental field sites. (2) Establish a multi-tier genomics-based analysis and ecological integration capability that links DOE JGI sequencing technology with need to understand functional abilities of soil microbial communities in an ecological setting, and to provide a platform that allows comparison of those processes across different terrestrial ecosystems. (3) Improve our basis for comparison of soil populations involved in carbon cycling and climate change response by expanding the functional genes and proteins we can use to detect and monitor these populations. Metagenomics can be usefully coupled with large, field-scale ecological studies that are focused on quantifying and modeling key processes.

In the next two years, six of the DOE's long-term free air CO_2 enrichment (FACE) and open top chamber (OTC) experiments will come to completion, offering an excellent opportunity to determine the effects of over ten year of elevated CO_2 treatment on below-ground ecosystem processes and the soil microbial communities responsible for those processes. The long-term FACE and OTC experiments encompass forest, scrubland, desert, and wetlands, allowing comparison of belowground responses of very different terrestrial ecosystems to elevated CO_2. Soil microbiota play critical roles in cycling carbon and nitrogen in terrestrial ecosystems, and their contributions have local, regional and global impacts on terrestrial carbon storage and cycling.

A multi-institutional, FACE soil metagenomics working group has been assembled to study soil microbial community structure across the FACE and OTC sites. Using a variety of comparative metagenomic sequencing approaches, we are addressing two questions: (a) Has long-term elevated CO_2 treatment affected the abundance and composition of soil microbiota in the soil? (b) Have the soil communities in different ecosystems responded similarly or in different ways to elevated CO_2? Three targeted metagenomic strategies are currently being used to address these questions across the DOE's FACE and OTC sites: (a) taxonomic profiling of the bacteria, archaea, and fungi in soils exposed to elevated or ambient CO_2 conditions, (b) sequencing suites of functional genes that are important in carbon and nitrogen cycling,

* Presenting author

(c) sequencing the soil fungal trasnscriptome in sites where we have evidence of the importance of fungal activities in response to elevated CO_2. We plan to also investigate seasonal responses to elevated CO_2, and the interactive effects of ozone and addition of soil nitrogen using these approaches. We will include a total community metagenomics approach in our future studies. We present here the initial results of rRNA profiling of bacterial and fungal communities across treatments at each of the FACE and OTC sites, and of high depth coverage of fungal rRNAs using 454 pyrosequencing for the aspen FACE site.

GTL

Flow Sorting and Whole Genome Amplification of Individual Microbes

Sébastien Rodrigue[1]* (s_rod@mit.edu), Rex R. Malmstrom,[1]* Aaron Berlin,[2] Matthew Henn,[2] and **Sallie W. Chisholm**[1,3] (chisholm@mit.edu)

[1]Dept. of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Mass.; [2]Broad Institute, Cambridge, Mass.; and [3]Dept. of Biology, Massachusetts Institute of Technology, Cambridge, Mass.

Project Goals: Develop a high throughput, ultra-clean pipeline for performing whole genome amplification on individual marine bacteria using flow cytometric cell sorting; determine the best technologies and strategies to sequence genome from individual microbial cells.

Genome sequencing of single cells is becoming a reality through the combination of whole genome amplification (WGA) and next-generation sequencing technologies. We have developed a high throughput, ultra-clean pipeline for performing WGA on individual marine bacteria using flow cytometric cell sorting and automated liquid handlers. The pipeline has already enabled WGA of several hundred individual bacterial during a single session. To validate the pipeline for sequencing, we performed WGA on individual cells from a culture of *Prochlorococcus* strain MED4, for which a reference genome had previously been assembled by Sanger sequencing. The resulting DNA was prepared for sequencing on the Roche 454-FLX and Illumina platforms. Both platforms yielded comparable coverage results, although 454 assemblies were superior. With 454-FLX, ~98% of the reference genome was covered at >1X after one sequencing run of single cell amplified DNA. A *de novo* assembly without a reference genome resulted in 523 contigs covering 84% of the reference *Prochlorococcus* MED4 genome, while unamplified DNA extracted from a culture containing billions of cells was assembled into only 7 contigs covering 99.85% of the reference genome. This difference between a single cell and billions of cells is due to uneven amplification of DNA from an individual cell; some genomic regions had >1000X coverage while others <5X. A comparison of four identically amplified *Prochlorococcus* MED4 single cell genomes revealed no patterns in amplification bias, in agreement with previous reports suggesting that uneven coverage is a stochastic process. To mitigate this uneven coverage, we developed a normalization procedure for 454 sequencing libraries that preferentially removes highly abundant sequences. This normalization procedure can reduce overall coverage variation by two orders of magnitude and dramatically improves *de novo* assembly of amplified single cell genomes.

GTL

Viruses Hijacking Cyanobacterial Carbon Metabolism

Luke R. Thompson[1]* (luket@mit.edu), JoAnne Stubbe,[1,2] and **Sallie W. Chisholm**[1,3]

[1]Depts. of Biology, [2]Chemistry, and [3]Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

http://chisholmlab.mit.edu/

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

Cyanophage infecting the marine cyanobacteria *Prochlorococcus* and *Synechococcus* carry several genes involved in the pentose phosphate pathway (PPP), a nighttime alternative to the Calvin cycle that generates NADPH and ribose. Oxidizing conditions in cyanobacteria at night, when photosystem I cannot generate NADPH, favor flux through the PPP, including the key enzyme transaldolase. Many cyanophage carry a transaldolase gene (*talC*) that differs markedly in structure from the host transaldolase (*talA*), suggesting that its acquisition and maintenance by cyanophage stems from functional differences with the host transaldolase. We have shown that the host enzyme is subject to oxidation in aerobic conditions, requiring reductant for full activity, whereas the phage transaldolase has no such effect. Site-directed mutagenesis of host transaldolase cysteines suggests that a mechanism independent of disulfide bond formation is responsible for this redox effect. We have recently found that another PPP-related gene, for the photosynthetic regulatory protein CP12, is also carried by many cyanophage. In cyanobacteria and other phototrophs, CP12 binds and deactivates two Calvin cycle enzymes under nighttime oxidizing conditions, promoting flux through the PPP. It therefore seems that cyanophage promote flux through the PPP by encoding not only enzymes but also a regulatory protein that inhibits the competing Calvin cycle. Phage infection of some cyanobacteria is known to lead to oxidizing conditions. Oxidative inactivation of host transaldolase and activation of phage CP12 may therefore be physiologically important, allowing cyanophage to produce NADPH and ribose for nucleotide biosynthesis and genome replication. Abundance patterns in phage genomes of *talC*, *cp12*, and two phage-encoded PPP dehydrogenases (*zwf* and *gnd*) are mirrored in environmental sequence databases, suggesting that the metabolic hijacking of cyanobacteria by cyanophage may be a globally important phenomenon.

Quantitative Proteomics of *Prochlorococcus*: Towards an Integrated View of Gene Expression and Cellular Stoichiometry

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

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

The marine cyanobacterium *Prochlorococcus* is the most abundant oxygenic photosynthetic organism on earth and a key component of oceanic carbon and nutrient cycling. With 13 sequenced genomes of cultured isolates and a wealth of observational data from natural environments, it is also a prime model system for understanding the molecular bases of microbial ecology and biogeochemistry. The small sizes of both genomes (~1700 genes) and cells (~0.6 micron) of *Prochlorococcus* also make tractable a relatively complete inventory of gene products and systems biology analysis of its streamlined metabolism. We have implemented a metabolic ^{15}N-labeling strategy to perform quantitative proteomic measurements, and find that it is comparable in accuracy and precision to mRNA-level measures of gene expression, including RT-qPCR, microarrays and RNA-sequencing.

In the oceans, cell division in *Prochlorococcus* populations is well-synchronized to the diel light-dark cycle, a phenomenon that can be reproduced in laboratory culture. Microarray experiments have demonstrated that most of the genome shows a substantial cyclicity of expression over this diel cell cycle. These results point to a 'just-in-time' mode of metabolism, where limited resources are apportioned in turn to various cellular processes in a coordinated fashion. We are extending these observations with quantitative proteomics, to explore differences in timing and magnitude between mRNA- and protein-level gene expression. The ^{15}N metabolic labeling strategy, in combination with accurate cell counts by flow cytometry, affords the ability to quantify proteins on a per-cell basis. Our aim is to develop a comprehensive picture of gene product stoichiometries (including both transcript and protein abundances) over the diel cell-division cycle of *Prochlorococcus*, which will help elucidate the molecular processes underlying microbial growth and carbon fixation in the global ocean.

Bacterioplankton Community Transcriptional Response to Environmental Perturbations

Jay McCarren* (mccarren@mit.edu), Rex Malmstrom, Yanmei Shi, Sebastien Rodrigue, **Sallie W. Chisholm**, and **Edward DeLong**

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

Environmental metagenomic techniques have revealed the vast genetic potential present in natural microbial communities. A remaining challenge is to examine gene expression in response to environmental variation in complex communities, with one goal being to determine the kinetics and impact of microbial response in gene expression on biogeochemical processes. Next generation sequencing has enabled cDNA profiling of microbial communities and we are developing this approach to track community responses to various environmental perturbations. Using this approach, we conducted field experiments to identify the genes, pathways, and organisms that respond to enrichments of either carbon substrates or nutrients in open ocean planktonic microbial communities. By tracking changes in whole community and *Prochlorococcus*-specific DNA and cDNA abundance throughout the experiment, we aim to identify gene suites that are up or down-regulated in response to these nutrient perturbations. Such investigations will also reveal which taxonomic components of the community respond to these amendments. Determining which genes allow which organisms to access new sources of nutrients and/or organic carbon substrates will shed light on carbon cycling dynamics in the environment. These transcriptomic experiments are providing new and specific insight into the functional dynamics and metabolic responses of natural microbial communities to environmental perturbation.

We are focusing both on whole community analyses (synecology), as well as on specific groups like *Prochlorococcus* (autecology). The autecological approach with *Prochlorococcus* has the advantage that these cyanobacteria can be separated from the bulk community by flow cytometric sorting. Using this approach, we conducted field experiments to identify the genes, pathways, and organisms that respond to enrichments of carbon substrates (high molecular weight dissolved organic material; HMW-DOM), and nutrients (inorganic PO_4 and mesopelagic water amendments) in open ocean planktonic microbial communities. By tracking changes in whole community and *Prochlorococcus*-specific DNA and cDNA abundance throughout the experiment, we have identified specific gene suites that are up or down-regulated in response to these nutrient perturbations. Phylogenetic analyses based on both community DNA and RNA abundances revealed that specific components of the bacterial community responded differentially to HMW-DOM additions. Gammaproteobacterial orders *Alteromondales, Thiotrichales* and *Xanthomonadales* responded positively to this DOM amendment, whereas *Prochlorococcus* and *Pelagibacter* relative cell numbers decreased. Preliminary

transcriptome analyses indicate that particular functional gene suites show parallel, functionally variable expression patterns in response to HMW-DOM enrichments. Detailed analyses of these transcriptomic experiments are providing new and specific insight into the functional dynamics and metabolic responses of natural microbial communities to environmental perturbation.

GTL

Grand Challenge in Membrane Biology: A Systems Biology Study of the Unicellular Diazotrophic Cyanobacterium *Cyanothece* sp. ATCC 51142

Michelle Liberton[1]* (mliberton@biology2.wustl.edu), Jana Stöckel[1], Alice C. Dohnalkova[2], Galya Orr[3], Jon M. Jacobs[3], Thanura Elvitigala[1], Eric A. Welsh[1], Hongtao Min[4], Jörg Toepel[4], Thomas O. Metz[3], Hans Scholten[3], Michael A. Kennedy[3], Garry W. Buchko[3], Nicole M. Koropatkin[5], Rajeev Aurora[6], Bijoy K. Ghosh[1], Teruo Ogawa[7], Jason E. McDermott[3], Katrina M. Waters[3], Christopher Oehmen[3], Gordon A. Anderson[3], Thomas J. Smith[5], Richard D. Smith[3], Louis A. Sherman[4], David W. Koppenaal[2,3], and **Himadri B. Pakrasi**[1] (pakrasi@wustl.edu)

[1]Dept. of Biology, Washington University, St. Louis, Mo.; [2]Environmental Molecular Sciences Laboratory, [3]Fundamental & Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Wash.; [4]Dept. of Biological Sciences, Purdue University, West Lafayette, Ind.; [5]Donald Danforth Plant Science Center, St. Louis, Mo.; [6]Dept. of Molecular Microbiology and Immunology, Saint Louis University School of Medicine, St. Louis, Mo.; and [7]National Laboratory of Plant Molecular Genetics, Institute of Plant Physiology and Ecology, Shanghai Institute for Biological Sciences, Shanghai, China

Project Goals: Membrane processes are critical to solving complex problems related to energy production, carbon sequestration, bioremediation, and other issues in energy and environmental science. Understanding how membrane processes fit into the overall cellular physiology and ecology requires a systems-level analysis of the genetics, biochemistry, and biophysics of membrane components and how molecular machines assemble, function, and disassemble as a function of time. *Cyanothece* is a marine cyanobacterium capable of oxygenic photosynthesis, nitrogen fixation, and heterotrophic growth in the dark. This unicellular organism has evolved an elaborate diurnal rhythm to temporally separate nitrogen fixation from oxygen production during oxygenic photosynthesis because of the oxygen-sensitive nature of the nitrogenase enzyme. The diurnal patterns of nitrogen fixation in the dark and photosynthesis in the light make *Cyanothece* a unique model organism for studying solar energy harvesting, carbon sequestration, metal acquisition, and hydrogen production. We sought to use systems biology approaches to determine the underlying cell signaling networks that govern the functions of cyanobacterial membranes and their components to accomplish this dramatic diurnal cycling. Our research brings together expertise in microbiology, biochemistry, proteomics and metabolomics, structural biology, imaging, and computational modeling and bioinformatics to achieve these objectives.

Cyanothece 51142 is a marine cyanobacterial strain notable for its ability to perform oxygenic photosynthesis and nitrogen fixation in the same single cell. These incompatible processes are temporally separated: photosynthesis is performed during the day and nitrogen fixation at night. As part of a complex diurnal cycle, these cells accumulate and subsequently mobilize storage inclusion bodies, specifically glycogen (carbon) and cyanophycin (nitrogen), making them natural biological batteries. In order to understand at a systems level how *Cyanothece* accomplishes these complex metabolic processes, we have undertaken a combination of ultrastructural, physiological, genomic, transcriptomic, proteomic, and metabolomic studies of this organism. High-resolution 3-D electron microscopy revealed that *Cyanothece* cells have a single extensive internal thylakoid membrane system. The genome of *Cyanothece* was sequenced (1) and found to contain a unique arrangement of one large circular chromosome, four small plasmids, and one linear chromosome. Global transcriptional analyses (2) uncovered 30% of genes with cyclic expression patterns and pinpointed a significant impact of nitrogen fixation on the diurnal cycle of different fundamental pathways. We have utilized the high-throughput accurate mass and time (AMT) tag approach to examine the proteome of *Cyanothece* 51142, and identified a total of 3,470 proteins with high confidence, which is approximately 65% of the predicted proteins based on the completely sequenced genome. These studies, as well as metabolite profiling, structural studies (3,4), and physiological measurements, when coupled with computational analysis and metabolic modeling, describe an organism in which tight control of cellular processes linked to storage of metabolic products for later usage is paramount for ecological success.

References

1. Welsh EA, *et al.* (2008) The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proc Natl Acad Sci U S A* 105: 15094-15099.
2. Stöckel J, *et al.* (2008) Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc Natl Acad Sci U S A* 105: 6156-6161.
3. Buchko GW, *et al.* (2008) Insights into the structural variation between pentapeptide repeat proteins - Crystal structure of Rfr23 from *Cyanothece* 51142. *Journal of Structural Biology* 162(1): 184-192.
4. Koropatkin N, *et al.* (2007) The structure of a cyanobacterial bicarbonate transport protein, CmpA. *Journal of Biological Chemistry* 282(4): 2606-2614.

This work is part of a Membrane Biology EMSL Scientific Grand Challenge project at the W. R. Wiley Environmental Molecular Science Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle.

GTL

Functional Analysis of Trace Nutrient Homeostasis in *Chlamydomonas* using Next Generation Sequencers

Sabeeha Merchant[1,2] (sabeeha@chem.ucla.edu), Madeli Castruita,[2] David Casero,[3] Janette Kropat,[2] Steven Karpowicz,[2] Shawn Cokus,[3] and Matteo Pellegrini[1,3]*

[1]The Institute for Genomics and Proteomics, [2]Dept. of Chemistry and Biochemistry, and [3]Dept. of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Calif.

Project Goals: Our goal is to use next generation sequencing technology to annotate algal genomes with the goal of increasing our understanding of algal biology.

Chlamydomonas, a chlorophyte alga in the green plant lineage, is a choice model organism for the study of chloroplast-based photosynthesis and cilia-based motility. The 121 Mb draft genome sequence, determined at 13X coverage is estimated to encode approximately 15,000 protein coding genes. Besides the pathways for oxygen evolving photosynthesis, dark respiration of acetate and hydrogen production, the gene repertoire reveals less-studied pathways for fermentative metabolism, suggestive of extraordinary metabolic flexibility. The operation of these bioenergetic pathways is dependent on metal cofactors like copper, iron, manganese and zinc, and accordingly these elements are essential nutrients for *Chlamydomonas*. In a copper-deficient environment, *Chlamydomonas* will modify the photosynthetic apparatus by substituting a heme protein—Cyt c_6—for an abundant copper protein—plastocyanin—that accounts for about half of the intracellular copper. This modification is viewed as a copper sparing mechanism and is dependent on a plant specific transcription factor CRR1. We have used digital gene expression (DGE) and RNA-Seq methodology to characterize the *Chlamydomonas* transcriptome under steady conditions of various degrees of copper-deficiency and in a bloom situation where cells deplete the copper as they divide. Both methods are quantitative and show excellent correlation with real time PCR indicative of a large dynamic range relative to microarrays. Direct vs. indirect responses to copper-deficiency are distinguished by comparison of the *crr1* transcriptome to that of wild-type cells. The analyses indicate previously unknown modifications of the photosynthetic apparatus and the potential for modification of bioenergetic pathways.

GTL

Whole-Genome Comparative Analyses Across an Environmental Gradient Reveal Surprisingly Rapid Bacterial Adaptation Mediated by Horizontal Gene Transfer

Alejandro Caro,[1] Jie Deng,[2] Jennifer Auchtung,[2] James Tiedje,[2] and **Konstantinos Konstantinidis**[1]*(kostas@ce.gatech.edu)

[1]School of Civil and Environmental Engineering and School of Biology, Georgia Institute of Technology, Atlanta, Ga.; and [2]Center for Microbial Ecology, Michigan State University, East Lansing, Mich.

Project Goals: Dr. Konstantinidis' group at Georgia Tech will conduct computational analyses of genomic, transcriptomic, proteomic, metabolomic and physiological data for *Shewanella* organisms to provide insight into the evolution, speciation and traits that determine niche optimization of these organisms. The work is going to be performed exclusively in-silico (computational), thus, no human or animal subjects or recombinant DNA etc. is involved. Dr. Konstantinidis' group will work cooperatively with Michigan State University and other members of the *Shewanella* Federation to achieve these goals. Georgia Tech's work is under Objective 4, Task 1 in the above project, entitled "Genetic and ecophysiological bases defining the core and diversification of *Shewanella* species".

How fast bacteria adapt to environmental fluctuations and to degrade new xenobiotic compounds remains a poorly understood issue of paramount importance for DOE's bioremediation efforts. One way that bacteria adapt is through changes in their genomic makeup, such as those caused by the exchange of genetic material among different bacterial strains. To provide new insights into these issues, we have sequenced and compared four strains of *Shewanella baltica*. These strains originated from four different depths of the stably stratified Baltic Sea, characterized by difference redox potentials and nutrient availability. The strains showed very similar evolutionary relatedness among each other, with their average genomic nucleotide identity being ~97%. Despite their comparable relatedness, the two strains isolated from more similar depths shared significantly more genomic islands compared to strains from different depths. The islands appeared to carry the ecologically important genes that determined strain's successful adaptation to the unique characteristics of the particular depth. Remarkably, the majority of these genes and an additional ~20% of the core genes (i.e. present in all genomes) showed 99.8-100% nucleotide identity between the two strains, suggesting that they had been horizontally exchanged between the strains in very recent evolutionary time. These results were validated against a larger collection of strains form the Baltic Sea using DNA-DNA microarrays. Collectively, our findings reveal that genomic adaptation could be very rapid, especially among spatially co-occurring strains, and advance our

understanding of population adaptation at redox interfaces, the most important environments for the transfer of pollutants relevant to DOE's missions.

GTL

A Laboratory Scientist Encounters Genome Sequences

F. William Studier* (studier@bnl.gov)

Biology Dept., Brookhaven National Laboratory, Upton, N.Y.

Project Goals: Improve genome annotation.

A consortium of scientists from Korea, France and the USA has determined the genome sequences of two B strains of *Escherichia coli*: REL606 is a strain used by Richard Lenski and colleagues for long-term studies of evolution in the laboratory; BL21(DE3) is a strain I constructed to be a host for expressing recombinant proteins under control of T7 RNA polymerase. Both are descended from the strain Delbrück and Luria named *E. coli* B in 1942 and which was adopted as the common host for the study of phages T1 to T7. Intensive study of laboratory strains derived from *E. coli* B and the closely related *E. coli* K-12, introduced by Tatum's 1944 work on biochemical mutants, is the basis for much of our current understanding of molecular genetics.

Comparison of the genome sequences of the two B strains has revealed the effects of the several different laboratory manipulations known to have occurred in the two lineages. Also revealed was an unknown misidentification in the literature that initially confused the analysis. The poster will summarize the differences between the two genome sequences, how they came about, and what they reveal about the effects of common laboratory practices.

GTL

Coupling Function to Phylogeny via Single-Cell Phenotyping

Michael C. Konopka[1]* (mkonopka@u.washington.edu), Sarah C. McQuaide,[2] Samuel Levine,[1] Ekaterina Latypova,[3] Tim J. Strovas,[4] Marina G. Kalyuzhnaya,[3] and **Mary E. Lidstrom**[1,3] (lidstrom@u.washington.edu)

[1]Dept. of Chemical Engineering, [2]Dept. of Electrical Engineering, [3]Dept. of Microbiology, and [4]Dept. of Bioengineering, University of Washington, Seattle, Wash.

Project Goals: One of the major challenges in understanding microbial community function in natural environments is linking genome-level sequences to function and role in the ecosystem. This project takes a new approach that couples a function-based live cell presorting step to single-cell analysis at both the physiological and genetic levels, using Lake Washington sediment and C1 cycling as the model. This approach will allow a culture-independent enrichment of live cells involved in specific functions, analysis of a variety of phenotypic capabilities at the single cell level, then targeting of those cells that test positive for specific functions for further culture-dependent and sequencing analysis. In a relatively rapid and high throughput manner, this system will identify cells with functions of interest and carry out a set of phenotypic tests on those cells. Subsequently, both culture-dependent and independent methods will be used to obtain sequence and phenotypic information that will couple function to genomics. The objectives of this project are 1) develop new technology for presorting functional populations and analyze them at the single cell level for both phenotypic and genomic parameters, and 2) apply this approach to populations from Lake Washington sediments to couple functional and genomic datasets at the single cell level.

Respiration represents the largest sink of organic matter in the biosphere, and is a fundamental component of global carbon flow. For decades, the balance between respiration and production has been considered an essential characteristic of ecosystems. From an environmental viewpoint, respiration is a key for understanding ecosystem functioning, structure and dynamics. We use respiration as a main physiological parameter to assess specific functions of yet uncultivable microbial population/individual cells from natural econiches. Here we present three approaches for detection and characterization of substrate-linked respiration: **1. Respiratory Detection System (RDS)** for bulk measurement of ecosystem function; **2. Respiration Response Imaging (RRI)** utilizing a fluorescent redox indicator for real-time detection of methylotrophic abilities of individual cells in a bulk system; **3. Microobservation Chamber** as a single cell analysis system for carrying out the physiological and genomic profiling of cells capable of respiring C_1-compounds.

1. Respiratory Detection System. To evaluate the potential of the RDS for environmental applications, the system was tested on Lake Washington sediment samples. For decades, this lake sediment has served as model system to study benthic carbon flux and microbial activities. It was demonstrated that aerobic reactions are restricted to a narrow zone of the lake sediment, and oxygen is depleted within the top centimeter of the sediment. We used the RDS to measure respiration rate in the top aerobic layer of the sediment cores. The minimum incubation time required for the rate calculation was 30-35 minutes (more then 40 times shorter than a routine incubation approach). The rate of oxygen consumption in the RDS system for fresh samples of the sediment was 1.4±0.05 mmol m^{-3} h^{-1}. The values obtained fall within the range previously reported for freshwater lakes (Giorgio and Williams, 2005). Overall, the RDS system provides precise measurements of oxygen decline over a short period of time (minutes, not hours), and thus shows great potential as a routine application in environmental studies.

2. Respiration Response Imaging. With RRI it is possible to detect the response of a natural population of microbial cells to an environmental change/stress at the very moment that it takes place at a single cell level. RedoxSensor green is a novel fluorogenic redox indicator dye (Molecular Probes, Invitrogen) that yields green fluorescence (488 nm excitation) after modification by bacterial reductases. Since respiration rate is proportional to fluorescence intensity, response to a new substrate or environmental change is easily detectable for individual cells from the entire population. We demonstrated using RRI to perform high throughput analysis of responses for environmental samples.

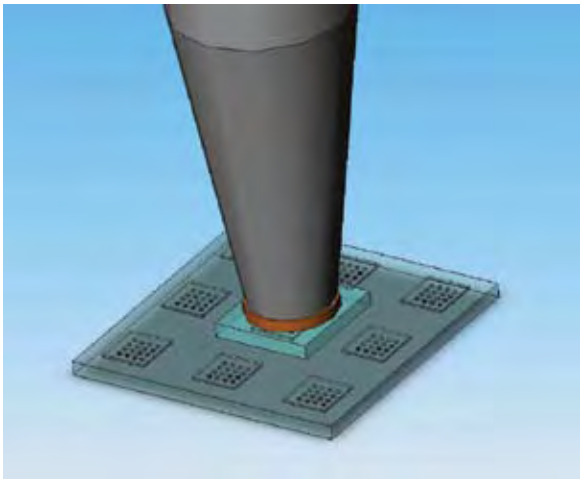


Figure 1. Glass chip containing arrays of microwells with piston sealing center array.

3. Microobservation Chamber. This system was based on an existing Microscale Life Sciences Center microwell-based platform consisting of a glass chip fabricated in-house with nine 4x4 arrays of etched microwells. Each microwell contains Pt-Porphyrin polystyrene beads as an oxygen sensor. The phosphorescent lifetime of the porphyrin sensor is inversely related to the oxygen concentration inside the well. Cells are seeded and the chip is placed inside a macro-well containing additional medium that bathes the chip. During an experiment, a piston with a flat glass tip is brought down over the top of the glass wells, coming into a pressure contact with the raised platform and diffusionally sealing each well from its neighbors (Fig. 1). Once the wells are sealed, lifetime measurements are made of each individual well in the array simultaneously and data are processed in real-time so that the oxygen consumption in each well can be monitored. After the oxygen consumption rate is detected inside the individual wells of an array, the piston is brought up and the wells are allowed to reoxygenate.

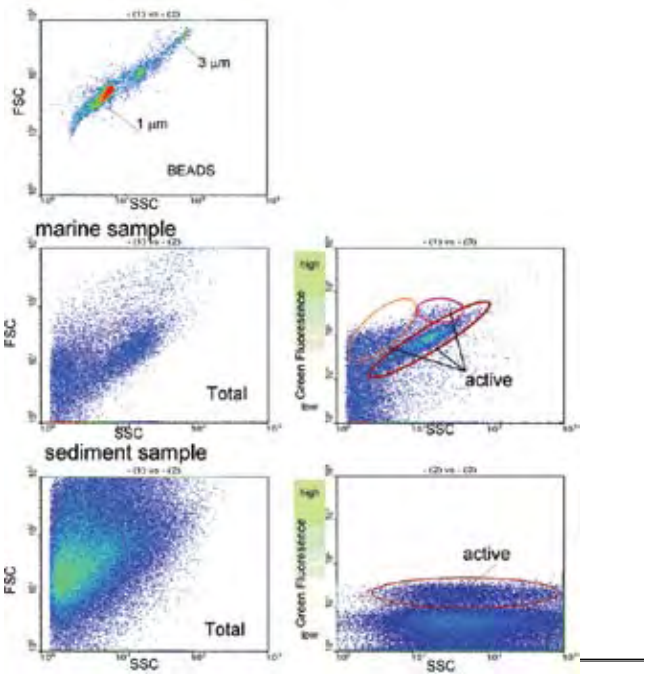


Figure 2. Flow cytometric analysis of cell populations from marine and sediment samples and stained with RSG.

We optimized the microwell design for bacteria and were easily able to detect O_2 consumption from individual microbial cells. This system was also successfully tested to measure the respiration rate of uncultured, live cells from the environment. We established a method for cell extraction and the examples of FC-plots of sediment and marine water samples are presented on Fig. 2. Microbial populations were separated by a CyFlowSpace flowcytometer/cell sorter (Partec); cells were then transferred onto the microchip platform and characterized in terms of respiration (O_2 consumption).

We demonstrated that these individual cells could be transferred to the 96 well plates using the Quixell Transfer micromanipulation system (Stoelting) and then used for whole-genome amplification (WGA) from single cells, and subsequent PCR analysis.

GTL

Proteomics Driven Analysis of Microbes, Plants and Microbial Communities

Mary S. Lipton[1]* (mary.lipton@pnl.gov), Stephen J. Callister,[1] Joshua E. Turse,[1] Thomas Taverner,[1] Margaret F. Romine,[1] Kim K. Hixson,[2] Samuel O. Purvine,[2] Angela D. Norbeck,[1] Matthew E. Monroe,[1] Xuixia Du,[1] Feng Yang,[1] Brain M. Ham,[1] Carrie D. Nicora,[1] Richard D. Smith,[1] and Jim K. Fredrickson[1]

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

Collaborators: Lucy Shapiro (Stanford University; Derek Lovley (University of Massachusetts); Steve Giovannoni (Oregon State University); Sam Kaplan (UT-Houston Medical School); Jerry Tuscan (ORNL); and *Shewanella* Federation

http://ober-proteomics.pnl.gov/

Project Goals: The "High-Throughput Proteomic Analyses of Microbial and Plant Systems" project exploits the proteomics pipeline at PNNL to address organism-specific scientific objectives developed in conjunction with biological experts for a number of different microbes and plants.

Significance: Characterization of biological systems using comprehensive global proteomic studies enhances scientific understanding through improved annotation of genomic sequences, 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.

Proteomic applications support DOE missions and science by exploiting microbial function for purposes of bioremediation, energy production, and carbon sequestration among other important areas. Inherent to exploiting microbial function or utilizing plants as biofuels is the detailed understanding of the physiology of the cell. These cellular functions are dictated by the proteins expressed in the cell, their localization and their modification state. The "High-Throughput Proteomic Analyses of Microbial and Plant Systems" project exploits the proteomics pipeline at PNNL 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 proteomics data for genome annotation of microbes and fungi, characterization of microbial communities, advances in the characterization of protein phosphorylation state, and the identification of new proteins important to photosynthesis, and the determination of protein localization in stem, root and leaf tissues of poplar.

The proteome can play an integral role in the identification of protein-encoding genes (CDS) in sequenced genomes as part of the initial annotation of a genome. For example, proteomics can be used to validate questionable gene identifications (e.g. encoding novel genes), identify CDS missed by automated gene calling algorithms, identify erroneous gene termini predictions, provide evidence for programmed frameshifting events that lead to alternative protein products, and provide evidence that sequencing mistakes are present. Up to now we have used proteomics data to enhance genome annotations after the initial annotation has been completed and publically released. In collaboration with JGI we are currently developing approaches to integrate proteomics data into the initial annotation pipeline. As part of this effort we have characterized numerous microbes and fungi that are targets of the GEBA genome sequencing project at the JGI to illustrate the utility of proteomic data in enhancing the quality of the annotation of genomic sequences and to allow subsequent cross-species comparative proteome analyses.

While comparative bacterial genomic studies commonly predict a set of genes indicative of common ancestry, experimental validation of the existence of this core genome requires extensive measurement and is typically not undertaken. Enabled by an extensive proteome database developed over six years, we have experimentally verified the expression of proteins predicted from genomic ortholog comparisons among 17 environmental and pathogenic bacteria. More exclusive relationships were observed among the expressed protein content of phenotypically related bacteria, which is indicative of the specific lifestyles associated with these organisms.

Populus is the fastest growing tree species in North America and has been identified as a potentially important crop species for converting plant biomass to liquid fuels. *Populus* species are broadly adapted to nearly all regions of the U.S., and hybrid clones have demonstrated 10 dry tons per acre productivity on a commercial scale. Still, improvements in growth rate, cell wall composition, drought tolerance, and pest resistance are required before this species reaches it potential as an energy crop. We have used proteomics technologies to map the protein expression patterns between root, leaf and stem tissues.

The *Rhodobacter sphaeroides* intracytoplasmic membrane (ICM) is an inducible membrane that is dedicated to the major events of bacterial photosynthesis, including harvesting light energy, separating primary charges, and transporting electrons. In this study, multichromatographic methods coupled with Fourier transform ion cyclotron resonance mass spectrometry and combined with subcellular fractionation, was used to test and prove the hypothesis that the photosynthetic membrane of *R. sphaeroides* 2.4.1 contains a significant number of heretofore unidentified proteins which are in addition to the integral membrane pigment-protein complexes previously discovered. These include light-harvesting complexes 1 and 2, the photochemical reaction center, and the cytochrome bc1 complex.

Our proteomic capabilities have been applied to characterize both the open ocean community in relation to *Peligabacter ubique* and the microbial community isolated from the termite (*Nasutitermes corniger*) hindgut. The proteome

characterization of these microbial communities presents a challenging application, and we are in the early stages of seeking to understand the ecology of these communities at the protein expression level and how this protein expression relates to the interaction of microbe with the environment and within the community. We show for *P. ubique* that a significant expression of the proteins involved in transport of metabolites and metals are indicative of the environmental metabolic requirements of this organism, and that this expression pattern is also represented in the larger sweater community.

Proteins regulate their function through expression levels and post-translational modifications, which can both be measured by proteomic analyses. Focusing on the characterization of the cell cycle in *C. crescentus*, we examined growth under carbon and nitrogen limitation conditions along with temporal resolution time courses to provide new insights on how this organism responds to its environment through genomic, proteomic, and ultimately morphologic strategies. We present results from the characterization of phosphorylation patterns of this organism, which revealed phosphorylation sites at threonine, serine, tyrosine and aspartate. Additionally, nine proteins observed to be up regulated, though these modifications are likely involved in elevated signaling processes associated with an adaptive response to the carbon starved growth environment.

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

*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.

GTL

High Throughput Comprehensive Quantitative Proteomics and Metabolomics for Genomics:GTL

Gordon A. Anderson* (gordon@pnl.gov) Ronald J. Moore, David J. Anderson, Kenneth J. Auberry, Mikhail E. Belov, Stephen J. Callister, Therese R.W. Clauss, Kim K. Hixson, Gary R. Kiebel, Mary S. Lipton, Thomas O. Metz, Matthew E. Monroe, Heather M. Mottaz, Carrie D. Nicora, Angela D. Norbeck, Daniel J. Orton, Ljiljana Paša-Tolić, Kostantinos Petritis, David C. Prior, Samuel O. Purvine, Yufeng Shen, Anil K. Shukla, Aleksey V. Tolmachev, Nikola Tolić, Karl Weitz, Rui Zhang, Rui Zhao, and **Richard D. Smith** (rds@pnl.gov)

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: 1) Continue to operate and expand high throughput proteomics measurement capabilities; 2) Increase overall data quality, proteome coverage, and quantitative measurement precision/accuracy, and provide data with statistically sound measures of quality; 3) Characterize and implement technology advances that specifically augment Aims 1 and 2; and 4) Develop and improve capabilities for managing, disseminating, and mining proteomics results in support of Genomics:GTL-supported projects.

New and expanded capabilities for quantitative high throughput proteomics and metabolomics are being developed and applied to achieve high levels of data quality that enable broad studies, e.g., of diverse microbial systems, communities, and potentially ecosystems. In concert with other measurements and information, these capabilities are increasingly providing the basis for new systems-level biological insights.

Advancing the understanding of microbial and bioenergy-related systems is at the heart of DOE's Genomics:GTL program. A key aspect for acquiring such biological understandings at a systems level is the ability to quantitatively measure the array of proteins (i.e., the proteome) for a particular system under many different conditions, as well as the abundances of a large (and often uncertain) range of metabolites and other cellular components. Among the basic challenges associated with making useful comprehensive proteomic measurements are: 1) identifying and quantifying large sets of proteins whose relative abundances typically span many orders of magnitude, and 2) making proteomics measurements that provide information on protein-proteins interactions and protein subcellular localization with sufficiently high throughput to enable practical systems biology approaches. Similar challenges exist for identifying and quantifying large sets of metabolites, in addition to even greater challenges for high throughput structural identification of metabolites due to broad structural diversity.

PNNL is addressing these issues by developing and applying new mass spectrometry-based measurement technologies in a high throughput environment to a range of collaborative GTL biological studies. Currently, our high throughput proteomics pipeline uses high resolution separations combined with mass spectrometry measurement capabilities that are integrated with advanced informatics tools amenable to the high data production rates and computational challenges associated with large-scale data comparisons. The accurate mass and time (AMT) tag strategy allows for both effective quantitative and high throughput peptide and intact protein-level analyses. Technological advancements related to sample processing and fractionation, as well as to the measurement platform itself have enabled an increasingly broad range of complex biological systems to be effectively addressed. Using the same mass spectrometry-based measurement platform, but with additional and complementary sample processing, fractionation and separation approaches, PNNL is also implementing complementary metabolomics approaches that aim to quantitatively define the broad range of other cellular constituents, and that are vital for understanding the function of biological systems.

Among the technological advancements for proteomics are new approaches that combine top-down and bottom-up measurements to extend quantitative proteome coverage to a large range of protein modification states. In another effort, a new fast separation LC-ion mobility-MS platform has been demonstrated to achieve high levels of data quality in conjunction with an order of magnitude increase in measurement throughput (see poster by R. D. Smith, et al.). These advances and new platform also directly benefit metabolomics, where measurements are complemented by the use of alternative separations and ionization modes in order to provide broad coverage of the diverse chemical universe of biological systems.

As part of these efforts, the supporting computational infrastructure at PNNL has also been expanded to incorporate a suite of search tools, data consolidation applications, and statistical relevance calculators, as well as visualization software for data interpretation. These informatics advances have enabled proteomic characterization of microbial communities from contaminated ground water and sediments, *Shewanella* strain comparisons, investigations of complex microbial response to oceanic seasonal cycling events, and more effective protein identifications, such as different modification states for *Caulobacter crecentis*. Moreover, these capabilities are also advancing GTL systems biology research; for example, proteomics measurements recently led to the discovery of active transport mechanisms involved in the uptake of nutrients in the oceans – a process previously thought to be governed by diffusion.

This poster will illustrate these technical advances, using specific applications to a range of Genomics:GTL projects as examples.

Acknowledgements
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.

GTL

A New Platform for Much Higher Throughput Comprehensive Quantitative Proteomics

Richard D. Smith* (rds@pnl.gov), Gordon A. Anderson, Erin Baker, Mikhail E. Belov, William F. Danielson III, Yehia M. Ibrahim, Ryan Kelly, Andrei V. Liyu, Matthew E. Monroe, Daniel J. Orton, Jason Page, David C. Prior, Keqi Tang, and Aleksey V. Tolmachev

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: 1) Continue to operate and expand high throughput proteomics measurement capabilities; 2) Increase overall data quality, proteome coverage, and quantitative measurement precision/accuracy, and provide data with statistically sound measures of quality; 3) Characterize and implement technology advances that specifically; augment Aims 1 and 2; and 4) Develop and improve capabilities for managing, disseminating, and mining proteomics results in support of Genomics:GTL-supported projects.

Significance: A new fast separation LC-ion mobility-MS platform for quantitative high throughput proteomics measurements has been developed that can achieve high levels of data quality in conjunction with an order of magnitude increase in measurement throughput. The new platform not only provides improved sensitivity and significantly speeds large-scale applications, but also lowers the cost of proteomics measurements. As a result, previously impractical studies of diverse microbial systems, communities, and ecosystems, among others are now possible.

Obtaining a systems level understanding of complex systems such as microbial communities, and ecosystems requires characterization of very large numbers of samples (e.g., involving many perturbations or spatially and/or temporally distinct samples). Although the proteomics measurement throughput attainable with LC-MS(/MS)-based approaches is much greater than with classical (e.g., 2D-PAGE-MS) approaches, it generally is grossly short of that needed for many systems biology applications. To address this shortcoming, we have been developing a new platform with greatly improved measurement throughput, sensitivity, robustness, and quantitative capability for proteomics measurements applicable to a range of GTL program interests.

Our proteomics platform encompasses fast capillary LC separations coupled via a greatly improved electrospray ionization (ESI) interface to an ion mobility spectrometer (IMS) that is interfaced to a high speed and broad dynamic range time-of-flight mass spectrometer (TOF MS). A fully automated capillary LC system incorporates high pressure LC pumps, an autosampler, and a 4-column nanoscale fluidics system. Each 10-cm-long nanocapillary LC column is operated at 10,000 psi. ESI-generated ions are accumulated in an electrodynamic ion funnel trap before being injected into an 84-cm-long IMS drift tube. To increase IMS-TOF MS sensitivity, we developed a novel multiplexing approach that increases the number of ion pulses that can be separated in a given time by >50-fold. Downstream of the IMS drift tube, diffusion dispersed ion packets are collimated by an electrodynamic ion funnel into a high performance orthogonal acceleration TOF MS. A high-performance data acquisition was developed that enables high mass accuracy high dynamic range measurements.

Initial evaluation of the LC-IMS-TOF MS system platform has shown significantly improved performance compared to the best existing proteomics platforms. In our initial evaluation, encoded MS signals were reconstructed, "deisotoped," and matched, using accurate mass and retention time information, against a reference database of peptides. The new platform was able to detect trace level peptides at high signal-to-noise ratios and an average peptide intensity coefficient of variance of ~9%. These results represent a significant improvement in data quality that is obtained in conjunction with the improvements in throughput.

Further improvements in performance from the use of an advanced ESI multiemitter source and ion funnel interface are expected to significantly build upon these initial results. In combination with improved informatics tools, this new proteomics capability is expected to enable previously impractical systems biology proteomics applications. Initial applications of the new platform are planned for the coming year.

Acknowledgements:
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.

GTL

Live-Cell Visualization of Tagged Bacterial Protein Dynamics and Turnover

Thomas Squier* (thomas.squier@pnl.gov), M. Uljana Mayer, and Yijia Xiong

Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: Newly synthesized MAPs built upon the cyanine dyes used for single molecule imaging offer the potential for multicolor measurements of protein localization and associations, and provide a path-forward for the high-throughput parallel characterization of protein-protein networks using a single tagging step and affinity technology. Because protein complexes can be released using simple reducing agents, low-affinity binding interactions can be captured, identified, and validated using the same MAPs. However, the robust utilization of MAPs requires the development of standard protocols that provide recipes and outline limitations regarding how MAPs can be used to image and purify protein complexes. Accordingly, we will focus on the application of existing MAPs and associated resins that we have synthesized, paying particular attention to the following deliverables. I. Demonstrate Utility of New Brighter MAPs (i.e., AsCy3) to Image Bacterial Proteins. II. Estabilish Ability of MAPs to Isolate Protein Complexes in Comparison with Established Tandem Affinity Purification Approaches. III. Benchmark Requirements of MAPs for Imaging and Protein Complex Measurements.

Biarsenical multiuse affinity probes (MAPs) provide an important tool to determine the abundance, location, size, binding interfaces, and function of tagged protein in living cells (1). Complementary measurements using MAPs immobilized on solid supports permit the isolation of intact protein complexes replete with low-affinity binding partners for complementary functional and structural measurements (2, 3). Our recent introduction of multiple tagging sequences and complementary MAPs with different colors and improved photostability (4-6), coupled with the ability to make switchable MAPs for single-molecule measurements that permit subdiffraction imaging (7, 8), offers a robust toolkit for high-throughput cellular measurements. However, while MAPs work well for the live-cell imaging of eukaryotic cells, their application in prokaryotes has been largely limited to the detection of highly overexpressed proteins in *E. coli* (9). Further, common protocols do not permit the stoichiometric labeling of tagged proteins, limiting their quantitative applications. To solve this latter hurdle in the application of MAPs in prokaryotes, we have systematically investigated methods that permit the specific labeling of tagged cytosolic proteins at near equimolar stoichiometries with minimal background signal from nonspecific labeling. Critical to our successful strategy was the expression of a tagged cyanofluorescent protein (CFP*) containing a C-terminus tag (i.e., CCPGCC), whose cytosolic location requires the delivery of the MAP (in this case FlAsH-EDT2) across the outer and inner membrane (Figure 1). Under optimal conditions, the tagged CFP* is selectively labeled; the small background signal for the empty-vector control is due to nonspecific labeling of endogenous bacterial proteins. Importantly, following expression of CFP*, which accounts for about 1% of cellular protein, nonspecific labeling is reduced. These results indicate that the tetracysteine tagging sequence on CFP successfully competes with nonspecific sites, resulting in specific labeling.

Figure 1. *Live-Cell Visualization of Cytosolic Proteins*. SDS-PAGE (left), live-cell images (center), and time-dependent labeling (right) showing *Shewanella oneidensis* MR-1 proteins following *in-vivo* labeling by FlAsH-EDT_2 (0.5 μM) for cells expressing an empty vector control (lanes 1 and 3; left panel or Control in center panel) or a CCPGCC tagged cyanofluorescent protein (CFP*) (lanes 2 and 4; left panel or Tagged Protein in center panel). SDS-PAGE (left) shows total protein stain (lanes 1 and 2) or fluorescence (lanes 3 and 4) for 40 μg total protein. Fluorescent images involve the direct excitation of FlAsH at 490 nm. Labeling Protocol. *Shewanella oneidensis* was cultured in LB medium at 30 °C; when the optical density reached 0.8, we simultaneously induced the expression of CFP* through the addition of arabinose (1 mM) and exposed cells to FlAsH (0.5 μM) in the presence of protamine (0.2 mg/mL) to enhance FlAsH uptake, and Disperse Blue 3 (20 μM) to reduce binding to nonspecific hydrophobic pockets. Labeling time is 2 hr. Cell Disruption. Cells were collected by centrifugation (5,000 x g), resuspended in PBS (50 mM NaH_2PO_4, 150 mM NaCl, pH 8.0), and cells were disrupted following the addition of lysozyme (5 mg/mL).

An example of the utilization of MAPs to access protein trafficking and turnover involved the metal reducing protein MtrC at the cell surface. These measurements took advantage of the probe CrAsH, whose net charge permits the selective labeling of tagged proteins located in the outer membrane. We find that MtrC is selectively labled, and that trafficking to the surface requires the presence of the type II secretory system involving gspG/gspD (**Figure 2**). Expressed MtrC is stable under suboxic conditions; however, upon shifting to aerobic culture conditions the MtrC

proteins are degraded ($k_{turnover}$ = 0.036 hr^{-1}). The preferential degradation of MtrC under aerobic conditions is consistent with the senstivity of this decaheme protein to oxidative damage, and suggests the presence of a protein degradative system that recognizes damaged proteins located on the extracellular face of the outer membrane.

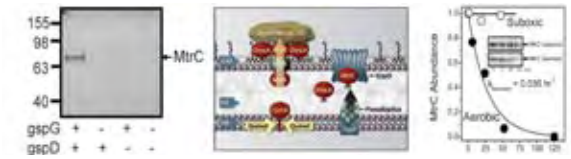


Figure 2. Specific Labeling of MtrC on cellular surface resolved using charged probe CrAsH (left panel), demonstrating that trafficking of MtrC involves type II secretion system (gspG and gspD) (center panel). Half-life of CrAsH-labeled MtrC is dependent on environmental conditions (right panel), consistent with role of MtrC in mediating electron transfer under suboxic conditions.

Conclusions:

- Equimolar stoichiometric labeling of tagged cytosolic proteins permits quantitative live cell measurements of bacterial protein dynamics.
- Selective labeling of outer membrane proteins permits measurements of trafficking and structures of membrane proteins.

References

1. Adams, S. R., Campbell, R. E., Gross, L. A., Martin, B. R., Walkup, G. K., Yao, Y., Llopis, J. and Tsien, R. Y. (2002) New biarsenical ligands and tetracysteine motifs for protein labeling in vitro and in vivo: synthesis and biological applications. *Journal of the American Chemical Society 124*, 6063-6076.
2. Mayer, M. U., Shi, L. and Squier, T. C. (2005) One-step, non-denaturing isolation of an RNA polymerase enzyme complex using an improved multi-use affinity probe resin. *Molecular Biosystems 1*, 53-56.
3. Verma, S., Xiong, Y., Mayer, M. U. and Squier, T. C. (2007) Remodeling of the bacterial RNA polymerase supramolecular complex in response to environmental conditions. *Biochemistry 46*, 3023-3035.
4. Cao, H., Xiong, Y., Wang, T., Chen, B., Squier, T. C. and Mayer, M. U. (2007) A red cy3-based biarsenical fluorescent probe targeted to a complementary binding peptide. *Journal of the American Chemical Society 129*, 8672-8673.
5. Chen, B., Cao, H., Yan, P., Mayer, M. U. and Squier, T. C. (2007) Identification of an orthogonal peptide binding motif for biarsenical multiuse affinity probes. *Bioconjugate chemistry 18*, 1259-1265.
6. Wang, T., Yan, P., Squier, T. C. and Mayer, M. U. (2007) Prospecting the proteome: identification of naturally occurring binding motifs for biarsenical probes. *Chembiochem 8*, 1937-1940.
7. Bates, M., Huang, B., Dempsey, G. T. and Zhuang, X. (2007) Multicolor Super-Resolution Imaging with Photo-Switchable Fluorescent Probes. *Science*.
8. Rust, M. J., Bates, M. and Zhuang, X. (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods 3*, 793-795.
9. Chen, B., Mayer, M. U., Markillie, L. M., Stenoien, D. L. and Squier, T. C. (2005) Dynamic motion of helix A in the amino-terminal domain of calmodulin is stabilized upon calcium activation. *Biochemistry 44*, 905-914.

GTL

Identification of All Engineering Interventions Leading to Targeted Overproductions

Sridhar Ranganathan* (sur152@psu.edu), Patrick F. Suthers, and **Costas D. Maranas** (costas@psu.edu)

Dept. of Chemical Engineering, Pennsylvania State University, University Park, Pa.

http://maranas.che.psu.edu

Project Goals: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

Existing computational methods (e.g., OptKnock, METAOPT, OptGene) for strain redesign generate engineering interventions only one at a time, limiting the array of choices presented to the biotechnologist. Metabolic flux data obtained through high-throughput experiments (e.g., MFA) are not used directly, instead the maximization of a surrogate of cellular fitness (i.e., max of biomass) is employed to generate metabolic flux predictions. To remedy these limitations, we present a computational framework that predicts all possible engineering strategies for the overproduction of targeted biochemicals. Instead of looking at engineering strategies one at a time, we identify engineering interventions by classifying reactions in the metabolic model depending upon whether their flux values must increase, decrease or become equal to zero to meet the pre-specified overproduction target. We hierarchically apply this classification rule for pairs, triples, quadruples, etc. of reactions. This leads to the identification of a sufficient and non-redundant set of fluxes that must change (i.e., MUST set) to meet a pre-specified overproduction target. Starting with this MUST we subsequently extract a minimal subset of fluxes that must actively be engineered (i.e., FORCE set) to ensure that all fluxes in the network are consistent with the overproduction objective.

We first demonstrated the developed methodology for succinate production in *Escherichia coli* using the most recent *in silico E. coli* model, *i*AF1260. The method not only recapitulates existing engineering strategies (e.g., overexpression of phosphoenolpyruvate carboxylase and elimination of the competing byproduct ethanol) but also reveals non-intuitive

ones that boost succinate production by performing coordinated changes on pathways distant from the last steps of succinate synthesis. Following this study, we addressed the overproduction of various biochemicals identified as promising biofuels.

GTL

Automated Construction of Genome-Scale Metabolic Models: Application to *Mycoplasma genitalium*

Patrick F. Suthers[1]* (suthers@engr.psu.edu), Vinay Satish Kumar,[2] and **Costas D. Maranas**[1] (costas@psu.edu)

[1]Dept. of Chemical Engineering and [2]Dept. of Industrial Engineering, Pennsylvania State University, University Park, Pa.

http://maranas.che.psu.edu

Project Goals: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

Currently, over 700 genomes (including eleven plant species) have been fully sequenced, however, only about 20 organism-specific genome-scale metabolic models have been constructed. It appears that metabolic model generation can only keep pace with about 1% of the fully sequenced genomes. In response to this flood of present and future genomic information, automated tools such as Pathway Tools and SimPheny™ have been developed that, using homology comparisons, allow for the automated generation of draft organism-specific metabolic models. This has shifted the burden towards curating the accuracy and completeness of the automatically generated, though draft, reconstructions. All of these reconstructions remain to some extent incomplete as manifested by the presence of unreachable metabolites and growth inconsistencies between model predictions and observed in vivo behavior.

In this work, we highlight the construction of *in silico* metabolic models for a minimal organism *Mycoplama genitalium* [1]. Key challenges for *M. genitalium* included estimation of biomass composition, handling of enzymes with broad specificities and the lack of a defined medium. Our computational tools were employed to identify and resolve connectivity gaps in the model (*i.e.*, GapFind and GapFill) as well as growth prediction inconsistencies with gene essentiality experimental data (*i.e.*, GrowMatch). The resulting curated model, *M. genitalium i*PS189 (262 reactions, 274 metabolites) was 87% accurate in recapitulating *in vivo* gene essentiality results for *M. genitalium*. Specifically, we found that the model correctly identified 149 out of a total of 171 essential genes (i.e., specificity of 87%) and 16 out of a total of 18 non-essential genes (i.e., sensitivity of 89%). This level of agreement not only meets, but exceeds thresholds for metabolic model quality put forth in the literature. Additional *in vivo* gene essentiality studies using a fully defined medium could usher a more accurate elucidation of the true metabolic capabilities of *M. genitalium*, as well as suggest improvements to the model.

We describe the application of iPS189 to drive the development of a defined growth medium. As noted above, gene essentiality experiments were performed using a non-defined medium, SP-4, which contains beef heart infusion, peptone supplemented with yeast extract and fetal bovine serum. The use of undefined medium can confound the characterization of gene essentiality, as the exact environmental conditions are not fully specified. Motivated by these shortcomings, we used the *i*PS189 metabolic model as a roadmap of the available transporters, metabolites and internal interconversions to seek out growth medium components necessary for biomass production.

Reference

1. Suthers, PF, MS Dasika, V Satish Kumar, G Denisov, JI Glass and CD Maranas (2008), *PLoS Comp Biol*, accepted

GTL

Synthetic Lethality Analysis Based on the *Escherichia coli* Metabolic Model

Alireza Zomorrodi* (auz107@psu.edu), Patrick F. Suthers, and **Costas D. Maranas** (costas@psu.edu)

Dept. of Chemical Engineering, Pennsylvania State University, University Park, Pa.

http://maranas.che.psu.edu

Project Goals: The general aims of this project are: (Aim 1) to generate integrated computational tools for the automated generation and curation of genome-scale models of metabolism for microbial and plant systems; (Aim 2) to automatically generate maps tracking the fate of labeled isotopes through genome-scale models; (Aim 3) to fully elucidate metabolic fluxes in genome-scale models using GC-MS or NMR data; (Aim 4) to leverage flux data for wild-type strain to identify all possible engineering strategies that lead to overproduction of a targeted product.

Essential genes are defined as genes whose deletion is lethal. By analogy, synthetic lethals (SL) refer to pairs of non-essential genes whose simultaneous deletion negates biomass formation. One can extend the concept of lethality by considering gene groups of increasing size where only the simultaneous elimination of all genes is lethal whereas individual gene deletions are not. Synthetic lethality results provide a birds-eye view of the redundant mechanisms avail-

able for redirecting metabolism and reveal complex patterns of gene utilization and interdependence. We developed a bilevel optimization-based procedure for the targeted enumeration of all multi-gene (and by extension multi-reaction) lethals for genome-scale metabolic models that outperforms exhaustive enumeration schemes by many orders of magnitude.

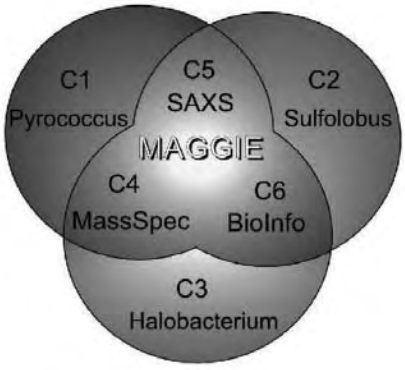
This proposed synthetic lethality analysis is applied to the *i*AF1260 model of *E. coli* K12 for aerobic growth on minimal glucose medium leading to the identification of all single, double, triple, quadruple and some higher-order SLs. The identified gene/reaction synthetic lethal pairs are phenotypically classified into two types: those yielding auxotroph strains that can be rescued through the supply of missing nutrients (i.e., amino acids or other compounds), and the ones lacking essential functionalities that cannot be restored by adding extra components to the growth medium. Graph representations of these synthetic lethals reveal a variety of motifs ranging from disjoint pairs, to hub-like stars, to *k*-connected sub-graphs. By analyzing the functional classifications of the genes involved in synthetic lethals (such as carbohydrate or nucleotide metabolism) we uncover trends in connectivity within and across COG functional classifications. We contrast the identified synthetic lethal predictions against experimental data and suggest a number of model refinement possibilities as a direct consequence of the obtained results.

We exhaustively assessed SL reaction triples and identified all reactions participating in at least one SL quadruples. The concept of degree of essentiality is introduced, defined as the lowest degree of a SL that a reaction participates, to unravel the contribution of each reaction in "buffering" cellular functionalities. We find that reactions in different COG classifications often involve very different degree of essentiality statistics. This study provides a complete analysis of gene essentiality and lethality for the latest *E. coli i*AF1260 and ushers the computational means for performing similar analyses for other genome-scale models. Furthermore, by exhaustively elucidating all model growth predictions in response to multiple gene knock-outs it provides a many-fold increase in the number of genetic perturbations that can be used to assess the performance of *in silic*o metabolic models.

Molecular Interactions and Protein Complexes

The MAGGIE Project

GTL



High Throughput Structural Characterization of Protein Complexes in Solution using Small Angle X-ray Scattering (SAXS) Combined with Mass Spectrometry (MS)

John A. Tainer[1,3]* (jat@scripps.edu), Greg L. Hura,[1] Steven M. Yannone,[1] Stephen R. Holbrook,[1] Jane Tanamachi,[1] Mike Adams,[2] Gary Siuzdak,[3] and Nitin S. Baliga[4]

[1]Life Science & Physical Biosciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]University of Georgia, Athens, Ga.; [3]The Scripps Research Institute, La Jolla, Calif.; and [4]Institute for Systems Biology, Seattle, Wash.

Project Goals: This is a Genomics:GTL proposal to support solution x-ray structures and other technologies to characterize Protein Complexes and Modified Proteins in microbes.

The GTL Project MAGGIE (Molecular Assemblies, Genes, and Genomics Integrated Efficiently) integrates teams at Lawrence Berkeley National Lab and the Advanced Light Source (ALS) with other researchers to characterize the Protein Complexes (PCs) and Modified Proteins (MPs) controlling microbial cell biology including stress and metabolic pathways relevant to bio-energy. Two key tools are Small Angle X-ray Scattering (SAXS) and Mass Spectrometry (MS). SAXS informs folding, unfolding, aggregation, shape, conformation, and assembly state in solution. SAXS resolution is limited to 50 to 10 Å resolution, but escapes the size limitations inherent in NMR and electron microscopy studies. We designed and built the SIBYLS synchrotron beamline (http://www.bl1231.als.lbl.gov) at the Advanced Light Source to interconvert between a SAXS and a crystallography endstation quickly allowing combined techniques. SAXS experiments and

results on native microbial biomass combined with MX and MS reveals the power of these combined techniques. SAXS results can determine correct molecular mechanisms for complexes involving conformational changes (Yamagata and Tainer, 2007). SAXS can identify extended or unstructured regions, and provide information on every sample at relatively high throughput (Putnam et al., 2007). The methods discussed provide the basis to examine molecular complexes and conformational changes relevant for accurate understanding, simulation, and prediction of mechanisms in structural cell biology and nanotechnology.

References

1. Putnam, C.D., Hammel, M., Hura, G. L., and Tainer, J. A. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution, *Quarterly Reviews in Biophysics* 40,191-285.
2. Yamagata, A, and Tainer, JA (2007). "Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism." *The EMBO Journal* 26, 878-890.

GTL

Define the Metalloproteome by Letting Metals Take the Lead: A Component of the MAGGIE Project

Aleksandar Cvetkovic[1]* (alcv@uga.edu), Angeli Lal Menon,[1] Farris L. Poole II,[1] Sarat Shanmukh,[1] Michael Thorgersen,[1] Joseph Scott,[1] Jeremy Praissman,[1] Ewa Kalisiak,[2] Sunia Trauger,[2] Steven M. Yannone,[3] John A. Tainer,[3] Gary Siuzdak,[2] and **Michael W.W. Adams**[1]

[1]Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens, Ga.; and [2]Center for Mass Spectrometry and [3]Dept. of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, Calif.

Project Goals: The overall goal of the MAGGIE project (Molecular Assemblies, Genes and Genomes Integrated Efficiently) is to provide robust GTL technologies and comprehensive characterization to efficiently couple gene sequence and genomic analyses with protein interactions and thereby elucidate functional relationships and pathways. The operational principle guiding MAGGIE objectives is that protein functional relationships involve interaction mosaics that self-assemble from independent protein pieces that are tuned by modifications and metabolites, including metals. The objective is therefore to comprehensively characterize the Protein Complexes (PCs) and Modified Proteins (MPs), including metalloproteins, underlying microbial cell biology.

Metal ion co-factors afford proteins with virtually unlimited catalytic potential, enable electron transfer reactions and are major determinants of protein stability. Consequently, metalloproteins play key roles in most if not all biological processes. However, it is not possible to predict the types of metal that an organism uses or the number and/or types of MPs encoded in its genome sequence because metal coordination sites are diverse and not easily recognized. The metalloproteome of an organism has yet to be defined. Directly determining the identity of MPs in native biomass utilizing a novel approach which focuses on the identification and purification of a metal instead of a protein which addresses some of these issues is described. We are currently using *Pyrococcus furiosus*, a hyperthermophilic archaeon that grows optimally at 100°C, as the model organism. Fractionation of native biomass close to 80°C below the optimal growth temperature using non-denaturing, chromatography techniques should provide both stable and dynamic protein complexes and metalloproteins for further characterization.

Liquid chromatography and high-throughput tandem mass spectrometry were used to separate and identify proteins, and metals were identified by inductively coupled plasma mass spectrometry (ICP-MS). After fractionation of a cytoplasmic extract from *P. furiosus*, a total of 249 metal peaks were identified, 154 of which appear to be unknown metalloproteins. These uncharacterized peaks included metals (uranium, lead, molybdenum, manganese and vanadium) that the organism was not previously known to utilize, as well as metals that were (iron, nickel, cobalt, tungsten and zinc). Similar chromatographic and metal analyses were performed for detergent-solubilized membranes of *P. furiosus* and for the cytoplasm of two other microorganisms, *Sulfolobus solfataricus* and *Escherichia coli* (grown on their conventional laboratory media). These revealed peaks of yet other types of metal, including tin, cadmium, zirconium, arsenic and thallium, all associated with as yet unknown proteins. Identification of these novel metalloproteins is underway and is being facilitated by the development of HPLC analyses with in-line ICP-MS. Our overall goal is to develop an analytical approach to define the metalloproteome of any organism under any growth condition.

These results indicate that metalloproteomes are much more extensive than previously recognized, and likely involve both biologically conventional and unanticipated metals with implications for a complete understanding of cell biology.

This research was funded by the Department of Energy (DE-FG0207ER64326) as part of the MAGGIE project.

GTL

The MAGGIE Project: Production and Isolation of Native and Recombinant Multiprotein Complexes and Modified Proteins from Hyperthermophilic *Sulfolobus solfataricus*

Robert Rambo,[1] Trent Northen,[1] Adam Barnebay,[1] Kenneth Stedman,[2] Michael W.W. Adams,[3] Gary Siuzdak,[4] Nitin S. Baliga,[5] Steven R. Holbrook,[1] John A. Tainer,[1,6] and **Steven M. Yannone**[1]* (SMYannone@lbl.gov)

[1]Dept. of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Center for Life in Extreme Environments, Portland State University, Portland, Wash.; [3]Dept. of Biochemistry and Molecular Biology, University of Georgia, Athens, Ga.; [4]Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, Calif.; [5]Institute for Systems Biology, Seattle, Wash.; and [6]Dept. of Biochemistry and Molecular Biology, The Scripps Research Institute, La Jolla, Calif.

Project Goals: As component 2 of the 6 component MAGGIE project our goals include 1) providing highly purified molecular machines from *Sulfolobus solfataricus* for physical characterization by SAXS (Component 5) and identification by MS/MS (Component 4), 2) Developing molecular biology in *Sulfolobus solfataricus* for recombinant protein expression in *Sulfolobus*, 3) Develop non-disruptive biochemical techniques to isolate and identify intact molecular machines from native biomass. 4) Develop methods to isolate and identify membrane protein assemblies and complexes. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

Dynamic protein-protein interactions are fundamental to most biological processes and essential for maintaining homeostasis within all living organisms. These interactions create dynamic and diverse functional networks essential to biological processes. Thus, a thorough understanding of these networks will be critical to engineering biological processes for DOE missions. The MAGGIE project was conceived, in part, as a response to the DOE GTL initiative to develop technologies to map the proteomes of model organisms. In this project we are exploiting unique characteristics of members of extremophilic Archaea to identify, isolate, and characterize multi-protein molecular machines. We have teamed expertise in mass spectrometry, systems biology, structural biology, biochemistry, and molecular biology to approach the challenges of mapping relatively simple proteomes. As part of the MAGGIE project, we are developing shuttle vectors for the extremophilic organism *Sulfolobus solfataricus* which has a growth optimum at 80°C and pH 3.0. We are using naturally occurring viral pathogens and plasmids to engineer shuttle vectors designed for recombinant protein tagging and expression in the native *Sulfolobus* background. The MAGGIE project has developed broadly applicable technologies for cellular deconstruction, isolation, and identification of intact molecular machines from native biomass. Our approach maintains a native environment throughout the characterization of the protein complexes and includes the complete partitioning of membrane proteins from the "soluble" material. Stable complexes are further characterized by small angle X-ray scattering (SAXS) at the Advanced Light Source at LBNL which provides the overall shape and size of the complex in solution. This work is directed toward developing rapid and efficient means to identify, isolate, and characterize molecular machines from any organism by integrating biochemical techniques with cutting-edge mass spectrometry, X-ray scattering and bioinformatics in a single approach. We are testing the idea that the hyperthermophilic nature of *Sulfolobus* will allow us to "thermally trap" protein complexes assembled at 80°C by isolating these complexes at room temperature. Ultimately, we aim to identify metabolic modules suitable to transfer specific metabolic processes between microbes to address specific DOE missions while developing generally applicable molecular and biophysical technologies for GTL.

GTL

Technologies for Robust and Rational Reengineering of Microbial Systems

W.L. Pang,[1]* T. Koide,[1] D.J. Reiss,[1] C. Bare,[1] K. Beer,[1] R. Bonneau,[2] S. Coesel,[1] M.T. Facciotti,[3] A. Kaur,[1] F.Y. Lo,[1] K. Masumura,[4] D. Miller,[1] M. Orellana,[1] M. Pan,[1] A.K. Schmid,[1] P. Shannon,[1] D. Tenenbaum,[1] P.T. Van,[1] K. Whitehead,[1] and **N.S. Baliga**[1] (nbaliga@systemsbiology.org)

[1]Institute for Systems Biology, Seattle, Wash.; [2]New York University, New York, N.Y.; [3]University of California, Davis, Calif.; and [4]National Institute for Health Sciences, Tokyo, Japan

Project Goals: Component 3: Bolster through high-end state-of-art systems approaches, developed specifically for the study of archaeal organisms, the comprehensive analysis of multi-protein complexes in DOE-relevant organisms.

Technologies to synthesize and transplant a complete genome into a cell have opened limitless potential to redesign organisms for complex specialized tasks. However, large scale reengineering of a biological circuit will require systems level optimization that will come from a deep understanding of operational relationships among all of a cell's constituent parts. We have developed systems approaches for the global deconstruction of transcriptional networks into statistically predictive models. We have also developed microfluidic platforms for reductionist analysis of sub-circuits within this global model to generate high-resolution kinetic expression

profiles that aid the development of fine-grained ODE/SDE based predictive simulations. The coupling of these two otherwise diametrically opposed approaches will provide avenues for the type of multiscale modeling necessary for systems reengineering.

GTL

Dynamic Assembly of Functional Transcriptional Complexes Inside Genes and Operons

David J. Reiss* (dreiss@systemsbiology.org), Tie Koide, W. Lee Pang, J. Christopher Bare, Marc T. Facciotti, Amy K. Schmid, Min Pan, Bruz Marzolf, Phu T. Van, Fang-Yin Lo, Abhishek Pratap, Eric W. Deutsch, Amelia Peterson, Dan Martin, and **Nitin S. Baliga** (nbaliga@systemsbiology.org)

Institute for Systems Biology, Seattle, Wash.

Project Goals: Component 3: Bolster through high-end state-of-art systems approaches, developed specifically for the study of archaeal organisms, the comprehensive analysis of multi-protein complexes in DOE-relevant organisms.

Transcription complexes assemble dynamically at specific sites on DNA to drive the controlled transcription of downstream genes. Despite knowledge of complex prokaryotic transcription mechanisms, generalized rules, such as the simplified organization of genes into operons with well-defined promoters and terminators, have played a significant role in systems analysis of regulatory logic in both bacteria and archaea. Through integrated analysis of transcriptome dynamics and protein-DNA interactions measured at high resolution throughout the *Halobacterium salinarum NRC-1* genome, we have identified widespread environment-dependent modulation of operon architectures, transcription initiation and termination inside coding sequences, and extensive overlap in 3' ends of transcripts for many convergently transcribed genes. We demonstrate that a significant fraction of these alternate transcriptional events correlate to binding locations of 11 transcription factors (TFs) inside operons and annotated genes – events often considered spurious or non-functional. This has illustrated the prevalence of overlapping genomic signals in archaeal transcription.

Protein Complex Analysis Project (PCAP)

GTL

Protein Complex Analysis Project (PCAP): Automated Particle Picking in Electron Microscopy Images

Pablo Arbelaez[1]* (arbelaez@eecs.berkeley.edu), Bong-Gyoon Han,[2]* (bghan@lbl.gov), Robert M. Glaeser,[2] and Jitendra Malik[1]

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

Project Goals: Develop a computer system capable of detecting automatically protein complexes in Electron Microscopy images. The method should be robust enough to process a large variety of particles and the quality of results should be comparable to human picking. Performance will be measured by the resolution of the final three dimensional reconstruction, as the main goal of the project is to propose a building block towards a fully automatic pipeline in Single Particle Analysis.

Particle picking is one of the main bottlenecks towards automating Single Particle Analysis. In this work, we propose a general framework for automatic detection of molecular structures in Electron Microscopy images. For this purpose, we formulate the problem as a visual pattern recognition task and use texture information as the main perceptual cue to differentiate particles from background. We follow a discriminative approach and train a non-linear classifier in order to precisely localize the particles in micrographs. Experiments conducted on several datasets of negatively stained specimens show that the reconstructions obtained from particles picked by our algorithm have resolution comparable with the ones obtained from human-picked particles.

GTL

A High Throughput Pipeline for Identifying Protein-Protein Interactions in *Desulfovibrio vulgaris* using Tandem-Affinity Purification

Swapnil Chhabra,[1] Gareth Butland[1]* (GPButland@lbl.gov), Dwayne Elias,[2] Veronica Fok,[1] Ramadevi Prathapam,[1] Thomas Juba,[2] John Marc Chandonia,[1] Ewa Witkowska,[3] **Mark Biggin**[1] (MDBiggin@lbl.gov), Terry Hazen,[1] Judy Wall,[2] and Jay Keasling[1,4]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]University of Missouri, Columbia, Mo.; [3]University of California, San Francisco, Calif.; and [4]University of California, Berkeley, Calif.

Project Goals: The broad goal of this subproject of PCAP is to study systems level biology of microbes of interest to DOE by developing a high throughput pipeline for facile genetic manipulation of these organisms. Specifically our goal is determine the genome-wide network of protein interactions, the abundance and spatial organization of the protein complexes within *Desulfovibrio vulgaris* Hildenborough (DvH), under normal laboratory and environmentally relevant stress conditions. These network inference data will be used to build comprehensive cellular models of metal reducing bacteria. The challenges faced in this project included the choice of a strictly anaerobic bacterium as the model soil bacterium, rudimentary genetic tools, and the effects of a high sulfide environment.

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 extend our functional genomics analyses of stress response pathways in *Desulfovibrio vulgaris*, we want to study the role of protein complexes in this sulfate reducing bacterium which has been found to exist in several DOE waste sites. Here we report the development of a technological platform for rapid identification of protein-protein interactions from a library of *D. vulgaris* mutant strains expressing tagged proteins. Our existing platform is based on the single cross-over chromosomal integration of tagged constructs generated in *E. coli* and we demonstrate the successful implementation of tagged-strain generation, verification and identification of interacting protein partners.

We demonstrate the strain generation process using automated software and hardware tools such as LIMS for automated sequence alignments and strain tracking, liquid handling systems for processing nucleic acids. We generated 363 tagged (STF/SPA) clones using the two-step TOPO-Gateway® cloning approach (Invitrogen) of which 278 were electroporated into competent *D. vulgaris* cells. We confirmed the single cross-over integration of 76 strains and the expression of affinity tagged fusion proteins using anti-FLAG IP-western blots and the verified strains were then subjected to TAP purification and MS based identification of interacting proteins. The TOPO-Gateway® mediated single-crossover approach works best for genes located at terminal ends of operons. In order to overcome these limitations, we have further developed the use of a double cross-over approach mediated through Sequence and Ligation Independent Cloning (SLIC). We will report our efforts towards adapting this method into a high throughput affinity tagged strain generation platform applicable to many species of interest to DOE.

GTL

Protein Complex Analysis Project (PCAP): Protein Complex Purification and Identification by "Tagless" Strategy

Ming Dong,[1]* Haichuan Liu,[2] Lee Yang,[1] Megan Choi,[1] Evelin D. Szakal,[2] Simon Allen,[2] Steven C. Hall,[2] Susan J. Fisher,[1,2] Gareth Butland,[1] Terry C. Hazen,[1] Jil T. Geller,[1] Mary E. Singer,[1] Peter Walian,[1] Bing Jap,[1] Jian Jin,[1] John-Marc Chandonia,[1] H. Ewa Witkowska,[2] and **Mark D. Biggin**[1] (mdbiggin@lbl.gov)

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

Project Goals: Water soluble protein complexes identified from 1% of the DvH proteome using the tagless strategy are presented. The data demonstrate that approximately 70% of polypeptides participate in multimeric assemblages. Multidimensional clustering methods are being developed to provide higher confidence for assigning putative heteromeric complexes based on similarity of polypeptide elution profiles. Many DvH homomers demonstrate unique stoichiometries that differ from those of their close orthologs.

Subproject B of the Protein Complex Analysis Project (PCAP) is developing several complementary high throughput pipelines to purify protein complexes from *D. vulgaris*, identify their polypeptide constituents by mass spectrometry and determine their stoichiometries. One of them is a "tagless" strategy for purifying and identifying water soluble protein complexes. The tagless strategy employs multidimensional separation of protein complexes followed by mass spectrometric monitoring of their composition. First, a crude cell extract (400 L culture) is separated under native conditions into a large number of fractions by an orthogonal four-step purification scheme. Then a rational sampling of ~30,000 of these fractions is made for MS analysis. Proteins are denatured, digested with trypsin, and then derivatized with primary amine-specific iTRAQ reagents. Samples are multiplexed (4- or 8-plexed, depending on the iTRAQ reagent used) and analyzed by LC MALDI MS/MS. In the course of MS/MS analysis, polypeptides are identified and, at the same time, their relative abundances are measured. As a consequence of the iTRAQ-based quantitation, elution profiles are generated for each polypeptide as it migrates through the column. We employ a clustering algorithm that

selects subsets of polypeptides that exhibit similar elution profiles to identify candidate heteromeric complexes. Polypeptides eluting without obvious partners at significantly higher MW than predicted from genome sequence data are categorized as putative homomers.

Over the last year, we have introduced further improvements to the tagless protocol, for example a miniaturized version of the 96-well PVDF multi-screen plate-based iTRAQ labeling procedure that lowers reagent cost by roughly a factor of ten. In addition, we have also developed an iterative MS/MS acquisition (IMMA) to facilitate greater sample analysis throughput by accelerating the rate limiting MS/MS analysis step (see additional poster).

To date, MS analysis of this pipeline has been performed on 395 fractions from 20 hydrophobic interaction columns that were in turn derived from 2 neighboring Mono Q column fractions. This represents around 1% of the total *D. vulgaris* water soluble protein fractionation space. Among 467 proteins identified, 263 were identified on the basis of 3 or more peptides, 343 on the basis of 2 peptides and the remaining 124 on the basis of 1 peptide. As judged by their elution profiles, the majority of polypeptides (~70%) migrated through a size exclusion column as multimeric assemblages. Seventeen heteromeric complexes were inferred on the basis of their homology to *E. coli* orthologs or a shared operon. In addition, many other polypeptides were found to cluster. Thus, to increase the level of confidence regarding putative interactions of polypeptides, a multidimensional clustering analysis algorithm aimed at comparing polypeptide elution profiles in all neighboring fractions is currently being developed. A total of 176 putative homomeric complexes were identified. One hundred of these have orthologs in *E. coli,* 55 of which have stoichiometries that differ between the two species. Independent analysis by electron microscopy of a number of these *DvH* homomeric complexes confirms that their stoichiometries differ not only from those of *E. coli* complexes but also in some cases from those of orthologs in more closely related species.

GTL

Protein Complex Analysis Project (PCAP): Localization of Multi-Protein Complexes through SNAP-Tag Labeling

Manfred Auer,[1,3]* **Mark D. Biggin**[1,3] (MDBiggin@lbl.gov), Gareth Butland,[1,3] Swapnil Chhabra,[1,3] Dwayne A. Elias[2,3] (eliasd@missouri.edu), Afolami Fagorala,[3] Terry Hazen,[1,3] Danielle Jorgens,[1,3] Dominique C. Joyner,[3] Thomas R. Juba,[2,3] Melissa Perez,[1,3] Jonathan P. Remis,[1,3] Andrew Tauscher,[1,3] and Judy D. Wall[2,3]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of Missouri, Columbia, Mo.; and [3]Virtual Institute for Microbial Stress and Survival, Berkeley, Calif.

http://vimss.lbl.gov/projects/pcap.html

Project Goals: The broad goal of this subproject of PCAP is to develop tag-based labeling approaches for high-throughput subcellular localization of proteins in microorganisms of interest to DOE. Our goal is to determine the abundance, the spatial organization and relative locations of proteins within individual, intact *Desulfovibrio vulgaris Hildenborough* (DvH) cells, as well as in microbial communities, under normal laboratory and environmentally relevant stress conditions. Further, we are attempting to correlate the intracellular abundance of proteins with toxic metal reduction activity. The challenges faced in this project included the choice of a strictly anaerobic bacterium as the model soil bacterium, rudimentary genetic tools, application of a fluorescent protein tag in an anaerobic bacterium for the first time and the effects of a high sulfide environment.

The key to an effective high-throughput approach has been the development of a functional genetic tagging approach. We chose to introduce tags onto single copies of the genes encoding the target proteins that were regulated by wild-type promoter sequences.

In initial attempts to visualize fluorescent tags, we encountered background labeling-problems with ReAsH labeling (Invitrogen) and, as such, have concentrated our efforts on SNAP-labeling (Covalys Technologies) because several SNAP-tagged *D. vulgaris* strains were available. Both the commercially available Gateway vectors and Sequence and Ligation Independent Cloning (SLiC) plasmids and methodologies were used to ligate the SNAP tag onto the carboxy terminal end of the gene of interest. Once the gene, tag and the direction of insertion into the plasmid were verified with PCR-based sequencing, the plasmid was electroporated into *D. vulgaris*. The resulting colonies were picked, checked for contamination, and the incorporation of the gene and tag into the *D. vulgaris* genome verified with PCR screening. If positive, cells putatively containing tagged constructs were grown and cell lysates subjected to an affinity/SNAP antibody column. The eluted protein was visualized on a denaturing gel and, if a protein band of the predicted molecular weight was detected with the SNAP antibody, this construct was considered confirmed. To date, it appears that the SLiC methodology for tagged gene plasmid production is cheaper, faster and has the advantage of ease of introduction of a promoter for expression of downstream genes in an operon.

The remaining challenges include a determination of the optimal screening and verification methods for the tagged construct. In some cases, Southern blots were performed and showed that the correct genetic structure had been achieved, yet no SNAP IP product could be detected. Conversely, cases existed where SNAP IP analysis was positive, but there was an absence of fluorescent labeling with either *in vivo* or *in vitro* labeling using SDS-PAGE analysis. In addition there has been heterogeneity of labeling among genetically identical bacteria, as well as our localization results in planktonic cells and biofilms using photoconversion approaches.

As a proof of principle for our fluorescent tag visualization, we have extended our approach to a small number of *E. coli*

constructs where the localization of the proteins are better established. These experiments will give us confidence that our procedures are working correctly. Once the labeling and photoconversion protocol are sufficiently robust, they can be applied to a growing number of DvH tagged strains under a variety of environmental conditions, including stress conditions. Once of the things we hope to learn is whether differences in protein expression levels, that are apparent in planktonic cells, may be the reason for cell-to-cell differences of metal reduction capability as seen in planktonic cells and in biofilms.

GTL

Protein Complex Analysis Project (PCAP): Survey of Large Protein Complexes in *Desulfovibrio vulgaris* Reveals Unexpected Structural Diversity

Bong-Gyoon Han[1]* (bghan@lbl.gov), Ming Dong,[1] **Mark D. Biggin**,[1] and Robert M. Glaeser[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: The aim of this component of PCAP is to develop high-throughput capabilities for determining the overall morphology and arrangement of subunits within large, biochemically purified multi-protein complexes of *Desulfovibrio vulgaris* Hildenborough. The quaternary structures of these multi-protein complexes will be used as templates to study proteomic network within whole cells in a native-like state by cryo-EM tomography.

Single-particle electron microscopy of multiprotein complexes purified by a tagless strategy has been used 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 k. The quaternary structures for 8 of the 16 complexes purified during this work were determined by single-particle reconstruction of negatively stained specimens, 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 multiprotein 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. It is further anticipated that imaging the spatial locations of such complexes, through the analysis of tomographic reconstructions (Downing et al.), may also be important for accurate computational modeling of such networks. While templates for some multi-protein complexes such as the ribosome or GroEL could be derived from previously determined structures, it is quite clear that single-particle electron microscopy should be used to establish the sizes and shapes of the actual complexes that exist in a new organism of interest. To not do so would be to risk searching for instances of a specific complex and finding none of them, simply because one had been searching with an invalid template.

GTL

Protein Complex Analysis Project (PCAP): Introduction of Iterative MS/MS Acquisition (IMMA) to the MALDI LC MS/MS Workflow To Enable High Throughput Protein Complex Identification using Tagless Strategy

Haichuan Liu,[1]* Lee Yang,[2] Nikita Khainovski,[2] Ming Dong,[2] Evelin D. Szakal,[1] Megan Choi,[2] Simon Allen,[1] Terry C. Hazen,[1] Jil T. Geller,[1] Mary E. Singer,[1] Peter Walian,[1] Bing Jap,[1] Steven C. Hall,[1] Susan J. Fisher,[1,2] H. Ewa Witkowska,[1] Jian Jin,[2] and **Mark D. Biggin**[2] (mdbiggin@lbl.gov)

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

Project Goals: Development and application of an intelligent method for MS/MS precursor ion selection is described in the context of the "tagless" strategy of protein complex identification. The Iterative MS/MS Acquisition (IMMA) algorithm allows for a step-wise execution of the MS/MS acquisition routine, where each step employs an automatically generated and updated exclusion list aimed at eliminating repetitive analyses of precursors derived from the previously identified polypeptides. A two-fold reduction in MS/MS acquisition time was achieved using this IMMA approach while increasing a number of polypeptides identified on the basis of 3 or more peptides by 25%.

The Protein Complex Analysis Project (PCAP) pursues two goals: (i) the identification of protein complexes in *Desulfovibrio vulgaris* Hildenborough (*DvH*) in order to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites and (ii) the development of workflows to enable high throughput analysis of protein-protein interactions in many other prokaryotes of interest to DOE. Protein complex identification is being performed by using two complementary approaches: the well established tandem affinity purification (TAP) and a novel "tagless"

strategy. The latter approach requires MS/MS-based protein identification from a large number of protein fractions (~30,000) that result from a four-step orthogonal separation of the entire *DvH* proteome under native conditions. Putative protein complexes are identified as a collection of polypeptides that co-elute in the final protein separation step, typically size exclusion chromatography (SEC). Polypeptide elution is monitored by iTRAQ reagent-based quantitation of changes in relative concentrations of each polypeptide as it migrates through the column.

Under the current experimental design, a single pass through the whole interactome employs ~5000 mass spectrometric samples, each containing an iTRAQ octaplex. The large number of MS/MS acquisitions required to fully characterize protein-protein interactions in a single organism calls for significant enhancements in throughput at various stages of analysis, of which MS/MS analysis is one of the main rate limiting steps.

There is a significant overlap in polypeptide content among neighboring fractions due to the limited resolution of protein chromatographic steps. For this reason, we have initially opted for LC MALDI MS/MS to execute the tagless strategy workflow because of the archival capability of this approach. Our plan has been to develop a precursor ion selection algorithm that would take advantage of the information that has already been gathered from previous MS/MS analysis. Here we introduce the concept of Iterative MS/MS Acquisition (IMMA) that forms the core of an overall software development effort geared towards substantially increasing throughput of the MS/MS data acquisition.

IMMA is an information-dependent precursor selection algorithm capable of eliminating the majority of redundant acquisitions triggered by precursors representing proteins already identified. This task is accomplished by generating exclusion lists of "undesired" precursors which, due to their relatively high intensity, would otherwise be selected for analysis by the data-dependent precursor selection software provided by the manufacturer of mass spectrometer. Two parameters characteristic of a peptide, molecular mass (m/z) and retention time (RT), are utilized to generate exclusion lists. Both parameters are modeled on the basis of the experimental data. Exclusion list candidates are limited to those peptides belonging to a proteotypic category, *i.e.,* those that are likely to be preferentially detected in the course of MALDI analysis (1). In addition, a peptide retention time (RT) prediction algorithm employs RT indices for amino acids and a set of modifications that we have frequently encountered in a training set. The RT prediction algorithm also allows for alignment of each newly performed LC run to a "virtual plate", *i.e.,* a predicted peptide retention time distribution that is built using training set data where a singular value decomposition algorithm is used to derive offsets for RT parameters. In addition to predicted proteotypic peptides, molecular ions already analyzed, satellite ions of high intensity precursors and contaminant ions are also excluded. In a typical experiment, one precursor is selected per spot at a time. The exclusion list is updated following each round of acquisition to reflect identification of new polypeptides and hence to eliminate an additional set of redundant precursors from further consideration. At present, execution of the IMMA algorithm allows for a two-fold reduction of acquisition time while using acceptance criteria of at least 3 confidently identified peptides per polypeptide. At the same time, the number of confidently identified polypeptides increased by 25%. Efforts are underway to extent the IMMA concept to multiple sets of LC MALDI runs, starting with all fractions from a single SEC column. In this scenario, the reference database of the already identified peptides will incorporate all species observed to date. Exclusion lists will be refined by using the exact rather than predicted retention times for species that were observed in earlier analyses. In addition, inclusion lists based upon peptides confidently identified in previous analyses will be executed to enable targeted monitoring of elution profiles of polypeptides identified in the same protein separation space.

We acknowledge members of Ruedi Aebersold's group for providing us with the "PeptideSieve" software for generating proteotypic peptides for DvH proteome according to their published algorithm: Mallick et al., (2007) Nat. Biotech. 25:125-131.

GTL

Protein Complex Analysis Project (PCAP): Isolation and Identification of Membrane Protein Complexes from *D. vulgaris*

Peter J. Walian[1]* (PJWalian@lbl.gov), Simon Allen,[2] Lucy Zeng,[1] Evelin Szakal,[2] Eric Johansen,[2] Steven C. Hall,[2] Susan J. Fisher,[1,2] Mary E. Singer,[1] Chansu Park,[1] Terry C. Hazen,[1] H. Ewa Witkowska,[2] **Mark D. Biggin**,[1] and Bing K. Jap[1]

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

Project Goals: As a component of the Protein Complex Analysis Project (PCAP) our goals are to 1) develop a pipeline for high-throughput isolation and identification of *Desulfovibrio vulgaris* membrane protein complexes and 2) use this methodology for identifying complexes isolated from wild type cells grown under normal conditions and to characterize changes in these complexes in response to environmentally relevant stressors.

This component of the Protein Complex Analysis Project (PCAP) has been focused on optimization of a processing pipeline to isolate membrane protein complexes from *D. vulgaris*, and identify their subunits by mass spectrometry. As part of PCAP's effort to model stress responses relevant to the detoxification of metal and radionuclide contaminated sites, we are applying a suite of the methods thus developed to catalog, as completely as possible, stable membrane protein complexes present in wild type cells grown under standard conditions as well as in the presence of stressors (*e.g.,* nitrate and sodium chloride). Our ultimate goal is to characterize changes in the relative abundance,

composition and stoichiometry of membrane protein complexes in cells with perturbed stress response pathways.

Membrane protein complexes are particularly challenging to purify and characterize. Largely due to the requirement for detergent solubilization, stable isolation of homogeneous intact membrane protein complexes in detergents typically requires separation conditions that are different from those used for water soluble proteins. Therefore, we are developing a novel "tagless" strategy specifically optimized for purifying membrane proteins and then identifying them by mass spectrometry.

In this approach, isolated *D. vulgaris* cell membranes are sequentially processed, initially using mild detergents suited for the solubilization of inner membrane proteins (such as Triton X-100 and C12E9) and then with a second more effective detergent (*e.g.,* octyl POE or octyl glucoside) to solubilize proteins of the outer membrane. Isolation of tagless complexes expressed in lower copy numbers is especially demanding and has driven us to process increasingly larger amounts of membrane (derived from cells obtained from over 100 liters of cell culture) in a given purification. To purify candidate complexes of the inner- and outer-membranes, ion exchange (IEX) and molecular sieve chromatography have been used. Fractions obtained from these procedures are further analyzed using SDS and blue native gel electrophoresis to isolate candidate complexes, obtain molecular weight estimates and to prepare complex subunit samples for mass spectrometry (MS) analysis.

To optimize chromatographic resolution and sample management under these conditions, proteins bound to ion exchange media are eluted using a fine step gradient profile. Following ion exchange chromatography, the collected fractions of each elution peak are concentrated and applied to a molecular sieve column. Molecular sieve column elution fractions are subjected to a combination of blue native and SDS PAGE. Proteins of potential complexes are extracted from the bands of blue native PAGE gels and directly transferred to a second dimension SDS PAGE. To improve upon extraction efficiency and the rate of sample production for mass spectrometry (MS), an alternative approach of gel-to-gel protein transfer has been employed in the past year. In this procedure, length-wise sections of blue-native PAGE gel lanes are being placed directly upon the stacking sections of denaturing gels and subjected to a second dimension of SDS PAGE. Protein bands or spots excised from these gels are subjected to in-gel digestion and analysis by liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS) and consequently identification by searching a custom *D. vulgaris* database using the Mascot search engine.

Using both native gel single-band extractions and two-dimensional techniques, over fifty membrane protein complexes have been identified. About one-fourth of these complexes are heteromeric.

We have recently begun preliminary analysis of *D. vulgaris* protein complexes from stressed cell cultures. Stressed cell conditions surveyed were stationary growth, elevated salt and elevated nitrate. While changes to proteins of the inner membrane, as assessed by SDS PAGE of IEX fractions, appeared to be relatively modest, changes to proteins of the outer membrane appear to be more pronounced. Changes to proteins of the outer membrane, in response to stressors, can be expected given that these proteins represent the first line of defense against environmental changes.

GTL

Protein Complex Analysis Project (PCAP): Reconstruction of Multiple Structural Conformations of Macromolecular Complexes Studied by Single-Particle Electron Microscopy

Maxim Shatsky[1,2]* (maxshats@compbio.berkeley.edu), Richard J. Hall,[2] Jitendra Malik,[2] and Steven E. Brenner[1,2]

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

In the hierarchy of biological visualization, 3D electron microscopy (EM) bridges the gap between the object sizes studied by X-ray crystallography and light microscopy. Single-particle EM is routinely used to resolve the three-dimensional structure of large macromolecular assemblies. As a part of LBNL's GTL PCAP project, single-particle EM molecular imaging is being applied to study protein complexes of *Desulfovibrio vulgaris*. These studies analyze stoichiometry and structural changes of the protein complexes under physiological and stress conditions. The structures obtained from EM experiments will serve as templates for identification and localization of the macromolecules in the images of the complete bacterial cells produced by the electron tomography. This structural information will facilitate the understanding of the functional roles of protein complexes in bioremediation studies by *D. vulgaris*.

One of the major challenges in single-particle EM is structural heterogeneity of the studied particles. Particles can adopt different conformations and can be found in assemblies with alternative quaternary-structure. Computational methods for image processing and three-dimensional structure determination play a crucial role in single-particle EM. Most commonly used computational approaches assume that the imaged particles have homogeneous shape and quaternary-structure. When this assumption is violated, the product of the three-dimensional reconstruction is one low resolution structure. Our aim is to improve and redesign various computational stages of particle reconstruction, taking into account the heterogeneity of the data.

We present a new computational method that reveals the existence of different conformational states of the studied macromolecular complexes. It is able to automatically classify the experimental images into homogeneous subsets which produce structurally different models. The method achieves high accuracy on synthetic data sets and shows

promising results on real data. In one test case, we successfully differentiated between the real experimental images of human translation initiation factor eIF3 and of eIF3 complexed with hepatitis C virus (HCV) IRES RNA. In another experiment, where we used experimental images of human RNA polymerase II, our method produced two models that show a substantial conformational flexibility of the protein complex. Our next goal is to apply this approach on various macromolecular complexes studied in PCAP. We expect that under some stress conditions protein complexes may be found in multiple structural configurations. Recognition of various structural conformations and, via electron tomography, their cell localization will aid in understanding of the *D. vulgaris* response to stress conditions related to bioremediation.

GTL

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris* : Data Management and Bioinformatics Subproject

Adam P. Arkin,[1,2] Steven E. Brenner,[1,2] Max Shatsky,[1,2] Ralph Santos,[1] Jason Baumohl,[1] Keith Keller,[1] and **John-Marc Chandonia**[1,2]* (jmchandonia@lbl.gov)

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

The Data Management and Bioinformatics component of the Protein Complex Analysis Project (PCAP) has two major goals: 1) to develop an information management infrastructure that is integrated with databases used by other projects within the Virtual Institute for Microbial Stress and Survival (VIMSS), and 2) to analyze data produced by the other PCAP subprojects together with other information from VIMSS to model stress responses relevant to the use of *D. vulgaris* and similar bacteria for bioremediation of metal and radionuclide contaminated sites.

We have developed a modular LIMS system to store data and metadata from the high-throughput experiments undertaken by the other PCAP subprojects. Each module of the LIMS corresponds to a step in the experimental pipeline. In the last year, we released WIST (Workflow Information Storage Toolkit), a template-based toolkit to facilitate rapid LIMS development. WIST allows LIMS programmers to design multi-step workflows using modular core components, which can be added and arranged through a simple, intuitive configuration and template mechanism. WIST uses the templates to create unified, web-based interfaces for data entry, browsing, and editing, and was used to build much of the current PCAP LIMS. WIST is available for download at http://vimss.sourceforge.net, under an open source license.

We have compared complexes identified in the PCAP tagless purification pipeline to complexes formed by orthologous proteins in other bacteria. We have discovered a surprisingly low degree of conservation in the stoichiometry of such complexes. This finding is consistent with the low degree of structural conservation that we observed in a smaller sampling of complexes studied using single particle EM (see the PCAP EM poster for details). We are currently analyzing results for outer membrane complexes under several stress conditions, and comparing data on the composition of complexes purified by the tagless approach to data on complexes isolated using affinity tags. We are also working with the Environmental Stress Pathway Project (ESPP) to integrate data on protein complexes into the MicrobesOnline website (http://microbesonline.org) for public dissemination.

GTL

Protein Complex Analysis Project (PCAP): Towards Localization of Functionality in *Desulfovibrio vulgaris* by Electron Microscopy

David A. Ball* (DABall@lbl.gov), Jonathan Remis, Manfred Auer, and **Kenneth H. Downing**

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: We aim to demonstrate the feasibility of using electron microscopy for high-throughput structural characterization of multi-subunit complexes in microbes. We will determine the spatial organization and relative locations of complexes within individual cells and map the relation between particular enzymes and other molecules and sites of metal reduction.

The anaerobic sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough (DvH) is used as a model organism for the study of environmental bioremediation of heavy metal and radionuclide contamination. DvH has the ability to efficiently reduce toxic heavy metals such as uranium and chromium and is of particular interest to DOE for use in high-risk metal contaminated sites. Electron microscopy is being used in several ways to gain insights on the molecular pathways involved in these and other processes.

Cryo-electron tomography and high-resolution single particle analysis (see abstract by BG Han et al., these proceedings) have been used to study the morphology of intact DvH cells and the larger protein complexes contained within them. During this specific study a number of surprising phenotypes and anomalies have been observed. The cytoplasm of DvH cells appears significantly more opaque than that of other bacterial species of similar size and doubling time, making it more difficult to identify molecular features. In cells that have been grown in sulfate up to and past cessation of growth, large internal "balls" have been observed, identified as sulfur deposits using Energy Dispersive X-ray Analysis. These large 'sulfurous balls' increase in

both size and number over time culminating in their occurrence across the total DvH population. Interestingly alongside the appearance of these electron dense sulfurous balls we find the emergence of a secondary population of rod-shaped lipid structures or "poles". These poles also appear to increase in number and regularity over time, are very beam sensitive and are present in two distinct length classes.

In addition to observing cells in their planktonic state, we have also examined DvH that were allowed to form a community (biofilm), which is more similar to their natural environmental habitat. Using high-pressure freezing and freeze-substitution or microwave-assisted processing, followed by resin-embedding, serial sectioning and electron tomography, we have identified a number of biofilm-specific structures including extracellular filaments and vesicle-like structures that assemble into clusters not unlike grapes on a vine, which are associated with extracellular metal deposition. Metal deposition was found to be non-uniform, with some cells appearing particularly active and others devoid of metal deposits. In addition to metal deposits near the bacterial cell surface, we have observed long-range sheet-like metal deposits that can extend for tens if not hundreds of microns. Based on their staining pattern we believe that these sheets contain a lipid core that is surrounded by metal deposits.

Current work is aimed at further characterizing localization of intra-and extracellular protein complexes that are involved in the cell's various metabolic pathways.

GTL

Protein Complex Analysis Project (PCAP): High Throughput Identification and Structural Characterization of Multi-Protein Complexes During Stress Response in *Desulfovibrio vulgaris:* Microbiology Subproject

Terry C. Hazen[1,4]* (tchazen@lbl.gov), Hoi-Ying Holman,[1,4] Jay Keasling,[1,2,4] Aindrila Mukhopadhyay,[1,4] Swapnil Chhabra,[1,4] Jil T. Geller,[1,4] Mary Singer,[1,4] Dominique Joyner,[1,4] Lauren Camp,[1,4] Tamas Torok,[1,4] Judy Wall,[3,4] Dwayne Elias,[3,4] and Mark D. Biggin[1,4]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]University of California, Berkeley, Calif.; [3]University of Missouri, Columbia, Mo.; and [4]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

The Microbiology Subproject of PCAP provides the relevant field experience to suggest the best direction for fundamental, but DOE relevant research as it relates to bioremediation and natural attenuation of metals and radionuclides at DOE contaminated sites. This project has built on techniques and facilities established by the Virtual Institute for Microbial Stress and Survival (VIMSS) for isolating, culturing, and characterizing *Desulfovibrio vulgaris*. The appropriate stressors for study have been identified and, using stress response pathway models from VIMSS, the relevance and feasibility for high throughput protein complex analyses is being assessed. We also produce all of the genetically engineered strains for PCAP. Three types of strains are being constructed: strains expressing affinity tagged proteins, those expressing fluorescent tags for sub-cellular localization, and knock out mutation strains that eliminate expression of a specific gene. We anticipate producing several hundred strains expressing affinity tagged proteins for complex isolation and EM labeling experiments by the other Subprojects. A much smaller number of knockout mutation strains are being produced to determine the effect of eliminating expression of components of putative stress response protein complexes. Both types of engineered strains are being generated using a two-step procedure that first integrates and then cures much of the recombinant DNA from the endogenous chromosomal location of the target gene. We are developing new counter selective markers for *D. vulgaris*. This procedure will; 1) allow multiple mutations to be introduced sequentially; 2) facilitate the construction of in-frame deletions; and 3) prevent polarity in operons. The Microbiology Subproject provides high throughput phenotyping of all engineered strains to determine if any show phenotypic changes. We also determine if the tagged proteins remain functional and that they do not significantly affect cell growth or behavior. The knockout mutations are tested in a comprehensive set of conditions to determine their ability to respond to stress. High throughput optimization of culturing and harvesting of wild type cells and all engineered strains are used to determine the optimal time points, best culture techniques, and best techniques for harvesting cultures using real-time analyses with synchrotron FTIR spectromicroscopy, and other methods. Finally, we are producing large quantities of cells under different conditions and harvesting techniques for optimal protein complex analyses. To insure the quality and reproducibility of all the biomass for protein complex analyses we use extreme levels of QA/QC on all biomass production. We expect to do as many as 10,000 growth curves and 300 phenotype microarrays annually and be producing biomass for 500-1000 strains per year by end of the project. Each biomass production for each strain and each environmental condition will require anywhere from 0.1 – 400 L of culture, and we expect more than 4,000 liters of culture will be prepared and harvested every year. The Microbiology Subproject is optimizing phenotyping and biomass production to enable the other Subprojects to complete the protein complex analyses at the highest throughput possible. Once the role of protein complexes has been established in the stress response pathway, we will verify the effect that the stress response has on reduction of metals and radionuclides relevant to DOE.

During the last year, the Microbiology Subproject produced biomass for multi-protein complex isolation and identification by mass spectrometry, and for imaging multi-protein complexes by electron microscopy. This year we have provided more than 2000 L of biomass consisting of more than 1,100 individual productions. Production volumes range from less than 10 mL of DvH wt, and mutants, for imaging

and development of high-throughput tagging and isolation methods, to 400 L of DvH wt for isolation of membrane protein complexes. We currently produce 100 L of DvH wt in five days, operating two 5 L fermenters in continuous flow mode in parallel. Extensive monitoring and assays are performed to ensure product quality and consistency, including continuous measurement of optical density and redox potential, and discrete sampling for AODCs, anionic composition (including organic acids), anaerobic and aerobic plating, total protein concentrations, PLFA and qPCR.

During the past year we have continued to evaluate three methods for genetic construction of tagged genes for TAP and visualization. These included the commercially available Gateway system and recombineering protocol against the recently developed SLiC (Sequence and Ligation independent Cloning) method. Further, we have continued to streamline the high-throughput pipeline for clone construction and implement QA/QC and verification protocols that are amenable to high-throughput. Early results suggest that the SLiC method is not only cheaper, faster and can tag any gene regardless of placement within the operon, but also yields a greater pipeline efficiency of up to 30% (verified construct/electroportation). Our current throughput level with ~30 electroporations per week yields ~5 confirmed constructs (16%) with the Gateway system with a turnover time of ~6 weeks from gene amplification through to a verified tag construct.

Using a commercial liquid handling system (Biomek NX) and in-house developed LIMS system for sequence analyses and sample tracking, we processed 576 genes from the *D. vulgaris* genome for tagged construct generation using the commercial TOPO-Gateway (R) strategy. Of these we were able to successfully confirm 360 constructs bearing the STF/SPA and SNAP tags. The tagged constructs were then fed into the electroporation pipeline for tagged strain construction based on the single cross-over chromosomal integration approach which is limited to genes located on terminal ends of their operons. For tagging genes located within operons using the double-recombination approach, we developed a novel scheme based on the Sequence and Ligation Independent (SLiC) cloning technique as detailed in the adjacent poster.

The Center for Molecular and Cellular Systems

GTL

An Integrative Strategy for the Determination of the Modular Structure of Functional Networks of *Rhodopseudomonas palustris*

William R. Cannon[2]* (William.Cannon@pnl.gov), Don Daly,[2] Mudita Singhal,[2] Lee Ann McCue,[2] Ronald Taylor,[2] Dale A. Pelletier,[1] Gregory B. Hurst,[1] Denise D. Schmoyer,[1] Jennifer L. Morrell-Falvey,[1] Brian S. Hooker,[2] Chongle Pan,[1] W. Hayes McDonald,[1,3] **Michelle V. Buchanan**,[1] and H. Steven Wiley[2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Pacific Northwest National Laboratory, Richland, Wash.; and [3]Vanderbilt University, Nashville, Tenn.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

Protein-protein interactions were used as the foundation for an integrative approach for determining the modular structure of *Rhodopseudomonas palustris* cellular networks. *R. palustris,* a metabolically versatile anoxygenic phototrophic bacterium, is the current target for the Genomics:GTL Center for Molecular and Cellular Systems (CMCS). Our analyses have focused on protein interactions observed under differing conditions for nitrogen metabolism in which either NH_4^+ (fixed nitrogen) or N_2 serve as the primary source of nitrogen.

We have developed an approach where multiple sources of information are integrated to provide a more comprehensive perspective of the cellular networks than can be provided by protein interactions alone. The interaction data are integrated with functional clues from operon structure and with manual and automated analysis of functional subsystems from molecular machines and cellular processes. Transcriptional regulatory elements are then overlaid on the protein-interaction maps to provide an integrative perspective on the cellular machines.

Using this approach, we have reconstructed a wide ranging, catalogue of protein complexes and interactions involved in a diverse and rich set of cellular processes including nitrogen fixation, electron transfer, photosynthesis, protein synthesis, ATP synthesis, central metabolism, fatty acid synthesis and uncharacterized processes.

Analysis of the functional network associated with nitrogen fixation identified a number of hubs and bottlenecks. Of the five proteins that are both hubs and bottlenecks, the three proteins that have functional annotations are all involved in electron transfer or synthesis of electron transfer proteins.

* Presenting author

The two unknown proteins, in addition, appear to also be related to electron transfer processes via their interaction partners. We have also compared the global interaction network and individual subnetworks with those inferred from protein interaction data gathered in other bacteria, such as *E. coli* and the nitrogen-fixing soil bacterium *Mesorhizobium loti*. Of the comparable interactions studied between *E. coli* and *R. palustris*, 20% are shared between the two species. Our comparative approach both improves the functional annotation of interaction networks of non-model species, and reveals fundamental architectural principles of the biochemical networks of microbes.

GTL

An Imaging-Based Assay with High Sensitivity for Confirming and Characterizing Protein Interactions

Jennifer L. Morrell-Falvey* (morrelljl1@ornl.gov), A. Nicole Edwards, Jason D. Fowlkes, Robert F. Standaert, Dale A. Pelletier, Mitchel J. Doktycz, and **Michelle V. Buchanan**

Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

Identifying and characterizing protein interactions are essential for understanding and modeling cellular networks. Several methods exist to assay protein interactions; however, none are known to provide both confirmation of protein interactions and simultaneous quantification of biophysical parameters (binding strengths and association/dissociation rates) *in vivo*. We are currently developing an approach that combines an imaging-based protein interaction assay with a fluorescence photobleaching and recovery technique (iFRAP), and computer simulations to provide a facile, general method for quantifying protein binding affinities *in vivo*. This protein interaction assay relies on the co-localization of two proteins of interest fused to DivIVA, a cell division protein from *Bacillus subtilis,* and green fluorescent protein (GFP). We have modified this imaging-based assay to facilitate high-throughput applications by constructing new vectors encoding N- and C-terminal DivIVA or GFP molecular tag fusions based on site-specific recombination technology and have determined the range of binding affinities that can be detected using this assay. The sensitivity of the assay was defined using a well-characterized protein interaction system involving the eukaryotic nuclear import receptor subunit, Importin α (Impα) and variant nuclear localization signals (NLS) representing a range of binding affinities. Using this system, we demonstrate that the modified co-localization assay is sensitive enough to detect protein interactions with K_d values that span over four orders of magnitude (1nM to 15μM). Moreover, the spatial confinement of the interacting proteins should also enable measurements of binding constants *in vivo* using iFRAP. Initial experiments demonstrate the anticipated decay of fluorescence at the cell poles indicative of a binding interaction. The statistical variance is reported as a function of K_d.

GTL

Functional Characterization of Protein Complexes and Cellular Systems in *Rhodopseudomonas palustris* using Stable Isotopic Labeling and Quantitative Proteomics

Chongle Pan[1]* (panc@ornl.gov), W. Judson Hervey IV,[1,2] Michael S. Allen,[1,3] Dale A. Pelletier,[1] Gregory B. Hurst,[1] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Graduate School of Genome Science and Technology, University of Tennessee-Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [3]University of North Texas, Denton, Tex.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

Responses to various types of stresses can be identified in microbial cells at the protein level using quantitative comparative proteomics. Proteins can be metabolically labeled with stable isotopes, such as ^{15}N and ^{13}C, by growing a microorganism of interest in a medium enriched in those stable isotopes. Stable isotope labeling enables large-scale accurate protein quantification by mass spectrometry, referred to as quantitative proteomics. Here, quantitative proteomics was used to characterize protein complexes and cellular systems in *Rhodopseudomonas palustris.*

Stable isotope labeling can be used to evaluate performance of affinity-tagging strategies for studies of protein-protein interactions both at the level of the protein complex, and at the level of the proteome. Affinity-purified protein complexes are often accompanied by background, non-specific proteins. In this study, authentic interacting proteins of a model complex, DNA-dependent RNA polymerase (RNAP), were successfully distinguished from artifactual co-isolating proteins by the isotopic differentiation of interactions as random or targeted (I-DIRT) method (A. J. Tackett et al. J. Proteome Res. 2005, 4 (5), 1752-1756). To investigate broader effects of bait protein production on bacterial metabolism, we compared proteomes from strains harboring the plasmid that encodes an affinity-tagged subunit (RpoA) of the RNAP complex with the corresponding wild-type strains using stable isotope metabolic labeling. Expression of plasmid-encoded bait protein significantly

induced the expression of several proteins involved in amino acid biosynthesis.

Cellular systems function not only via a physical interaction network but also within a regulatory network. In results from the Center for Molecular and Cellular Systems, observations of protein-protein interactions among a putative anti-σ factor RPA4224, an extracytoplasmic function (ECF) σ factor RPA4225, and the predicted response regulator RPA4223 led us to study this system further. We characterized a global stress regulon controlled by RPA4225 in *R. palustris* using quantitative proteomics. Changes in expression of several genes resulting from overproduction of RPA4225 were further verified by quantitative PCR. Furthermore, most of the strongly up-regulated proteins revealed a conserved binding motif, which we also found in the promoters of over 150 genes, including general stress proteins. These data suggest that RPA4225 controls a global stress regulon that may be conserved among several members of α-Proteobacteria.

These studies showcased the biological insights one can obtain using stable isotope labeling and quantitative proteomics. In addition, new methods based on stable isotope labeling are under development at the Center for Molecular and Cellular Systems for protein absolute quantification and complex dissociation kinetics.

GTL

Protein-Protein Interactions in *Rhodopseudomonas palustris* at the Genomics:GTL Center for Molecular and Cellular Systems

Dale A. Pelletier[1]* (pelletierda@ornl.gov), Kevin K. Anderson,[2] William R. Cannon,[2] Don S. Daly,[2] Brian S. Hooker,[2] H. Steven Wiley,[2] Lee Ann McCue,[2] Chongle Pan,[1] Manesh B. Shah,[1] W. Hayes McDonald,[1] Keiji G. Asano,[1] Gregory B. Hurst,[1] Denise D. Schmoyer,[1] Jenny L. Morrell-Falvey,[1] Mitchel J. Doktycz,[1] Sheryl A. Martin,[1] Mudita Singhal,[2] Ronald C. Taylor,[2] and **Michelle V. Buchanan**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [2]Pacific Northwest National Laboratory, Richland, Wash.

Project Goals: The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species.

The overall goal of the Center for Molecular and Cellular Systems (CMCS) is to provide a capability for generating high quality protein-protein interaction data from a variety of energy- and environment-relevant microbial species. The CMCS approach combines expression of affinity tagged proteins, affinity purification of interacting proteins, and tandem mass spectrometric identification of these proteins (Pelletier et al., *J. Proteome Research* **2008**, *7*, 3319-3328).

We have recently completed the characterization of soluble protein-protein interactions in *Rhodopseudomas palustris*. These results are the first large-scale protein-protein interaction results of this type in an organism other than a model system such as *E. coli* or yeast. The protein-protein interactions from the metabolically versatile *R. palustris* provide insights into microbial processes of high relevance to DOE missions, including the ability to produce hydrogen, to degrade lignin monomers, to perform photosynthesis, and to fix nitrogen.

As of early December 2008, nearly 1200 *R. palustris* genes have been cloned as Gateway entry vectors, and approximately 1060 expression clones for a dual affinity tag (6-His/V5) have been produced. Over 800 affinity-tagged bait proteins have been expressed, affinity purified, and subjected to mass spectrometry (MS) analysis to identify interacting proteins. Criteria for choosing these bait proteins from among the >4800 in the *R. palustris* predicted proteome included predicted location in the cytosol, and previous detection at medium to high levels in proteomics measurements (VerBerkmoes et al., *J. Proteome Research* 2006, *5*, 287-298). Quantitative estimates of confidence in putative bait-prey interactions identified from the MS analysis are obtained using the statistical tool BePro3 (Sharp et al., *J. Proteome Research* 2007, 6, 3788-3795). From thousands of putative interactions, a few hundred survive as candidates for further study, based on preliminarily chosen BePro3 threshold values. Integration with additional data, including comparative genomics, transcriptomics, operon structure, and regulatory networks, that aid in interpretation and functional annotation of novel interactions is described in the poster "An Integrative Strategy for the Determination of the Modular Structure of Functional Networks of *Rhodopseudomonas palustris*" by Cannon *et al*. Results of the protein-protein interaction survey in *R. palustris* are available through the publicly accessible Microbial Protein-Protein Interaction Database (MiPPI.ornl.gov).

Validation of Genome Sequence Annotation

GTL

Toward a High Throughput Functional Annotation Pipeline for Fungal Glycoside Hydrolases

Scott E. Baker[1]* (scott.baker@pnl.gov), Ellen A. Panisko,[1] Deanna Auberry,[1] Beth Hofstad,[1] Jon K. Magnuson,[1] Barbara Robbertse[2], Adrian Tsang[3], and Frank Collart[4]

[1]Chemical and Biological Process Development Group, Pacific Northwest National Laboratory, Richland, Wash.; [2]Oregon State University, Corvallis, Ore.; [3]Concordia University, Montreal, Canada; and [4]Argonne National Laboratory, Argonne, Ill.

Project Goals: Our goal is to characterize glycoside hydrolases using an approach that is based on bioinformatic and proteomic analyses, high throughput protein expression and enzymatic activity assays.

Glycoside hydrolases (GHs) from fungi are a key component of the biological disassembly of cellulosic biomass into its component sugar monomers. In the DOE vision of the biorefinery, these sugars are subsequently converted to fuels and chemical products by microbial bioprocesses. As the number of fungal genome sequences increases, so too does the number of GHs that lack any functional characterization. We have initiated an approach to GH characterization that is based on bioinformatic and proteomic analyses, high throughput protein expression and enzymatic activity assays. In order to assess the diversity of an important class of GHs we performed a phylogenetic analysis of GH family 7. Our analysis indicates that GH7s have been lost or duplicated numerous times across the fungal kingdom. Additionally, our results show that encoded within the genome of a many individual fungal species, there is a tremendous range of amino acid sequence diversity that may be indicative of the variety of environmental growth conditions encountered by these organisms.

We are characterizing GHs within a variety of individual fungal species. Initial studies are bioinformatic and proteomic in nature. These are followed by high throughput heterologous protein production and enzymatic activity assays of individual GHs. *Aspergillus niger*, a filamentous ascomycete fungus, was the initial organism chosen for GH characterization in our pipeline. Current organism targets undergoing proteomic analyses include the basidiomycete brownrot, *Postia placenta*, a close relative of *A. niger*, *Aspergillus carbonarius*, the zygomycete, *Phycomyces blakesleeanus* and a thermophilic ascomycete fungus, *Thielavia terrestris*.

GTL

Annotation of the *Clostridium phytofermentans* ORFome by Proteogenomic Mapping

Andrew Tolonen*

Dept. of Genetics, Harvard Medical School, Boston, Mass.

Project Goals: The goal of this project is to analyze the *Clostridium phytofermentans* ORFome by building a proteogenomic map from data obtained by mass specrometry.

Clostridium phytofermentans is gram positive anaerobe that efficiently converts the two most abundant constituents of plant feedstocks, cellulose and hemicellulose, to ethanol and hydrogen. The breakdown of cellulosic biomass is accomplished by a diverse set of saccharolytic enzymes; the resulting hexoses and pentoses are then fermented, with ethanol as the primary product. *C. phytofermentans* is thus a model system for the direct conversion of cellulosic feedstocks to biofuels. The *C. phytofermentans* genome was sequenced by the DOE JGI. The genome is 4.8 Mb and contains 3926 candidate protein-encoding gene models. In this project we validated and corrected the JGI gene annotations by building a proteogenomic map from data obtained by mass specrometry.

A proteogenomic map is an application of proteomics to genome annotation: novel proteins not included in the standard annotation are identified, the boundaries of annotated ORFs are defined, and ORFs that are considered 'hypothetical' based upon computational gene prediction are verified. Proteogenomic mapping is an ideal complement to computational genome annotation because mass spectrometry provides direct, molecular evidence about which ORFs are translated into proteins. In this study, we analyzed the proteome of *C. phytofermentans* growing on different carbon sources: glucose, cellulose, and hemicellulose using an LTQ_FT Ultra Hybrid mass spectrometer. Greater than 10,000 unique peptides were mapped to the *C. phytofermentans* genome to create a proteogenomic map. This presentation will compare features of the computational genome annotation of *C. phytofermentans* to the genome annotation from the proteogenomic map.

GTL

Protein Functional Assignment Using a Fluorescence-Based Thermal Shift Assay

Ashley M. Frank[1]* (afrank@anl.gov), Sarah E. Giuliani,[1] William Studier,[2] and **Frank Collart**[1]

[1]Biosciences Division, Argonne National Laboratory, Lemont, Ill.; and [2]Biology Division, Brookhaven National Laboratory, Upton, N.Y.

Project Goals: Generate specific functional assignments for selected protein families such as ABC transporters, transcription factors and two component systems. Assemble an integrated data set that enables identification of sequence motifs associated with biological function of specific protein families.

We are developing *in vitro* methods for functional characterization of proteins that can be used to interrogate proteins involved in cellular metabolic, sensory and response pathways. Our approach for the identification of bound ligands and assignment of function uses a fluorescence-based thermal shift (FTS) assay for high-throughput screening of protein-ligand interactions. The FTS approach is a target-independent assay that uses a fluorescent dye to monitor protein unfolding. This assay uses a commercially available real-time PCR instrument, where thermal melting curves of the protein/ligand combinations can be screened in a 96- or 384-well plate format.

Our initial study focused on solute-binding proteins in the bacterial ABC transporter family. These transporters are essential membrane transport components in many organisms and transport a diverse range of ligands, but a specific functional role has not been assigned for a majority of these proteins in most organisms. The assay was validated with a set of six proteins with known binding specificity and was consistently able to map proteins with their known binding ligands. The assay also identified additional candidate binding ligands for several of the amino acid binding proteins in the validation set.

We extended this approach to additional targets and demonstrated the ability of the FTS assay to unambiguously identify preferential binding for several homologs of amino acid binding proteins with known specificity and to functionally annotate a protein of unknown binding specificity. The targets in this evaluation set had various degrees of experimental functional characterization (Table 1) but none of the individual proteins were represented as structures in the PDB. In all cases, the assay predicted ligand assignments that were consistent with ligand assignments inferred by other experimental approaches.

The FTS approach was also applied to a set of targets from *S. oneidensi*s which are clustered into two COGs (Table 1). Five of the *Shewanella* proteins are categorized into COG083ET, a cluster which contains amino acid binding and signal transduction proteins. These five proteins were annotated as periplasmic binding proteins but only one (NP_716672) was specifically annotated as an amino acid-binding protein. The remaining protein is grouped in COG2998H, a cluster with the description of "ABC-type tungstate transport system, permease component," but is annotated as a hypothetical protein (NCBI database). The *S. oneidensis* target selected from COG2998H was strongly stabilized by tungstate and showed some stabilization with molybdate. This protein has a signal peptide and is part of an operon encoding an ABC-type transporter and ATPase. Specific binding to arginine was detected for a different protein (NP_716672) which was originally annotated as an amino acid binding protein (Table 1).

Table 1. FTS Assay Results for Solute-Binding ABC-type Transporter Proteins Varying in Degree of Experimental Functional Annotation

Description	Source Organism	Accession #	Tm (°C) (no Ligand)	Binding Ligand(s)	Tm Shift 1000uM (°C)	Tm Shift 200uM (°C)	Tm Shift 20uM (°C)
[illegible]	*Escherichia coli*	NP_415391	59.0	ARG	9.0	5.0	0.0
DL-methionine transporter subunit	*Escherichia coli*	NP_414725	55.5	MET	6.0	4.0	1.0
Lysine, arginine, ornithine-binding protein	*Salmonella enterica*	NP_460611	48.0	ARG ORN LYS	9.0 7.0 6.0	6.0 4.0 4.0	0.0 1.0 0.0
COG083ET protein	*Shewanella oneidensis* MR-1	NP_719077	62.5	ARG	11.0	7.0	1.0
COG083ET protein	*Shewanella oneidensis* MR-1	NP_717123	59.5	More	—	—	—
COG083ET protein	*Shewanella oneidensis* MR-1	NP_717497	55.5	More	—	—	—
COG083ET protein	*Shewanella oneidensis* MR-1	NP_719550	57.5	More	—	—	—
COG083ET protein	*Shewanella oneidensis* MR-1	NP_719553	53.0	More	—	—	—
COG2998H protein	*Shewanella oneidensis* MR-1	NP_720224	51.5	Tungstate Molybdate	22.0 4.0	—	—

The assay is implemented in a microwell plate format and provides a rapid approach to validate an anticipated function or to screen proteins of unknown function. The ABC-type transporters family is ubiquitous and transports a variety of biological compounds, but the current annotation of the ligand binding proteins is limited to mostly generic descriptions of function. The results illustrate the feasibility of the FTS assay to improve the functional annotation of binding proteins associated with ABC-type transporters and suggest this approach that can also be extended to other proteins families.

GTL

Assignment of Enzymatic Function for Core Metabolic Enzymes

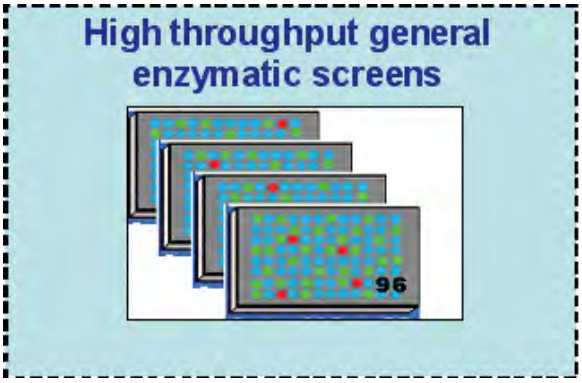
Elizabeth Landorf,[1]* Vincent Lu,[1] Gopi Podila,[2] Michael Proudfoot,[3] Alexander Yakunin,[3] and **Frank Collart**[1] (fcollart@anl.gov)

[1]Biosciences Division, Argonne National Laboratory, Lemont, Ill.; [2]Dept. of Biological Sciences, University of Alabama, Huntsville, Ala.; and [3]Banting and Best Dept. of Medical Research, University of Toronto, Toronto, Canada

Project Goals: Benchmark the utility of general enzymatic screens for the improvement of functional assignments. Evaluate the utility for specific assignment of function and the overall impact on the annotation set.

GTL

With over 800 genomes with complete sequences currently available in public databases and thousands of genome sequence projects in progress, there's a pressing need to effectively annotate genomic sequences quickly and accurately for functional activity. The main objective of this proposal is to experimentally annotate (assign a biochemical function) a large group of conserved hypothetical proteins using high throughput protein production and enzymatic screening methods. In the first stage of the project we have used the general enzymatic screens to functionally map a subset of the conserved hypothetical proteins from *Shewanella oneidensis* which contains ~800 members. To date, we have cloned ~500 cloned targets and 313 of these targets had levels of expression and solubility. Over 300 proteins were purified under Ni-affinity chromatography and ~200 of this set have been delivered to the University of Toronto for enzymatic screening. The screening data from this initial set suggest high throughput enzymatic screens have general utility for the identification of a functional category. Approximately 10-20% of the screened targets showed some activity using the current library of chemical compounds.



A directed screening strategy process was applied to a subset of proteins that demonstrated thioesterase, esterase or HAD hydrolase activity. For the targets that tested positive for thioesterase substrates, the directed screening was able to identify specific activity profiles for acetyl CoA substrates. Similar activity and biochemical profiles could be generated for other target families identified in the general screening process.

These studies benchmark the utility of enzymatic screening for functional annotation for proteins of unknown function. These results also suggest that bioinformatic prescreening and restriction of the generalized screens to specific ligand categories can improve the specificity and overall impact of the functional assignment. This approach will provide a foundation to assess the capabilities for specific functional assignments for a substantial number of unknown prokaryotic and eukaryotic proteins.

Quality Improvement Process for JGI-ORNL Microbial Annotation Pipeline

Miriam Land[1]* (landml@ornl.gov), Loren Hauser,[1] Janet Chang,[1] Cynthia Jefferies,[1] Gwo-Liang Chen,[1] Frank Larimer,[1] Natalia Mikhailova,[2] Natalia Ivanova,[2] Athanasios Lykidis,[2] Galina Ovchinnikova,[2] Amrita Pati,[2] Nikos Kyrpides,[2] and **Bob Cottingham**[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [2]DOE Joint Genome Institute, Walnut Creek, Calif.

Project Goals: With the extension of high-throughput sequencing to include microbial genomes, there has been a rapid expansion of microbial genomic data, requiring the development of comprehensive automated tools to provide in-depth annotation to keep pace with the expanding microbial dataset. We continue to develop tools for comparative multiple genome analysis that provide automated, regularly updated, comprehensive annotation of microbial genomes using consistent methodology for gene calling and feature recognition. An automated analysis pipeline provides annotation for the microbial sequencing projects being carried out at the JGI. Multiple gene-finders are used to construct a candidate gene model set. The conceptual translations of these gene models are used to generate similarity search results and protein family relationships; from these results a metabolic framework is constructed and functional roles are assigned. Simple repeats, complex repeats, tRNA genes and other structural RNA genes are also identified. Annotation summaries are made available through the JGI Microbial Sequencing web site; in addition, results are integrated into interactive display schemes at ORNL. Comprehensive representation of microbial genomes requires deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, repressor and activator binding site calling, transcription terminators, and other functional elements. Sensor development is continuing to enhance access to these features. Linkage and integration of the gene/protein/function catalog to phylogenomic, structural, proteomic, transcriptional, and metabolic profiles are being developed. The expanding set of microbial genomes comprises an extensive resource for comparative genomics: new tools continue to be developed for rapid exploration of gene and operon phylogeny, regulatory networking, and functional proteomics.

The US DOE Joint Genome Institute (JGI) performs high-throughput sequencing and annotation of microbial genomes through the DOE Microbial Genome Program (MGP). The world-wide rate of sequencing is resulting in a rapid expansion of microbial genomic data, which requires the development of comprehensive automated tools to provide in-depth annotation which can keep pace with the expanding microbial dataset.

JGI-ORNL annotates the microbial genomes which are sequenced by JGI. We have and continue to develop tools for genome analysis that provide automated annotation of microbial genomes using consistent methodologies. One such tool is our new gene caller, Prodigal, which has its own abstract.

Unfortunately, all automated methods have inherent errors because they are based on comparisons to existing data sources which vary in quality and applicability. We have implemented processes to monitor and improve the quality of the automated annotation while making the process fully automated. This will reduce the sources of error and will provide users with qualifiers on the annotation.

The JGI annotation process includes a quality control (QC) step. The annotation is reviewed for pseudo genes, missed genes, and potential changes in start codons. Patterns identified in the QC process are fed back into the initial annotation process to improve the quality of gene predictions. Continual improvement is also made to the QC process to make it both efficient and effective.

Inference flags have been added to the GenBank annotation. The inference and notes provide users with basis for determining the source and quality of the annotation.

ORNL has created a baseline database consisting of representative sequences from a taxonomically diverse group of organisms. This representative group of DNA sequences was initially annotated in early 2008. As the underlying databases change and tools are upgraded, the baseline database is reannotated and an impact analysis is done. This ensures that changes to the automated annotation have a positive effect on the outcome. In addition, genomes originally annotated by JGI-ORNL are being reannotated to improve the quality of their annotation available to the user community.

The JGI is made up of affiliates from a number of national laboratories including Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and the Hudson Alpha Genome Center.

GTL

Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

Valérie de Crécy-Lagard[1]* (vcrecy@ufl.edu) and Andrew D. Hanson[2] (adha@ufl.edu)

[1]Dept. of Microbiology and Cell Science and [2]Dept. of Horticultural Sciences, University of Florida, Gainesville, Fla.

Identifying the function of every gene in all sequenced organisms is a central challenge of the post-genomic era. We are submerged in genomic, transcriptomic, and proteomic data but the functions of about half (range 20 to 60%) of the genes in any given organism are still unknown. Our goal is to predict and experimentally verify the *in-vivo* function of proteins that lack homologs of known function ('unknown' protein families) and that are highly conserved between prokaryotes and plants. Our approach combines the extensive post-genomic resources of the plant field with the use of comparative genomic tools made possible by the availability of thousands of sequenced microbial genomes. This is an integrative approach to predict gene function whose early phase is computer-assisted, and whose later phases incorporate intellectual input from expert plant and microbial biochemists. It allows bridging of the gap between automated homology-driven annotations and the classical gene discovery efforts driven mainly by experimentalists. We have already analyzed 350 protein families and linked 50 of them to particular metabolic areas. In a second round of analysis we have predicted a testable function for 20 protein families and started experimental validation for 15 as summarized in Table 1. This has already led to experimentally supported annotations for >600 individual genes (Fig. 1A) and has the potential to yield such annotations for at least 3,000 more (Fig. 1B). From this work, which represents around half our planned analysis, it is clear that specific predictions are readily possible, and that these could never have been made without this type of integrative effort. It is also clear that this is a cost-effective way to assign function to unknown proteins. The predictions and status of the experimental validations will be available on a website in early 2009.

Fig. 1. Genes annotated so far (A) and potential for future annotations (B)

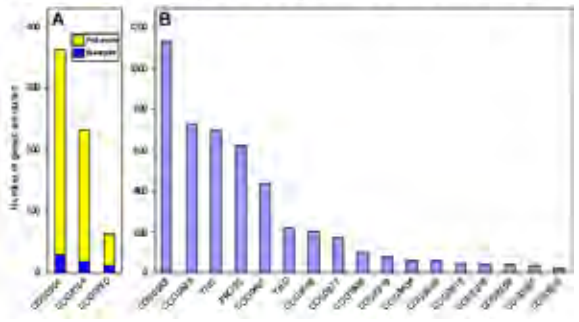


Table 1. Progress in predicting and validating functions of unknown protein families

Unknown protein family (*Arabidopsis* AGI codes)	Predicted function	Experimental validation (genes tested)	Genes annotated Actual [Potential]
COG2154 (At1g29810, At5g51110)	PCD with accessory role in Moco metabolism	Completed1 (20)	233
COG0354, YgfZ (At4g12130, At1g60990)	Fe-S cluster synthesis/ repair	In progress (7)	365
COG0720 (In Heterokonts, not higher plants)	Replaces FolB (& FolQ) in folate biosynthesis	In progress (9)	64
COG0212 (At1g76730)	Alternative 5-formyltetrahydrofolate cycloligase	In progress (5)	[46]
COG3404 (At2g20830)	Replaces 5-formyltetrahydrofolate cycloligase	In progress (6)	[66]
COG1836 (At1g78620)	Phytyl phosphate kinase	In progress (2)	[100]
COG3533 (At5g12960, At5g12950)	Glycosylhydrolase, plant cell wall breakdown	In progress (2)	[60]
COG3146 (At2g23390)	Pterin-dependent hydroxylase	In progress (1)	[200]
COG4319 (At3g09250, At4g10925)	Folate metabolism	In progress (1)	[81]
COG0009, YrdC (At5g60590)	t6A biosynthesis	In progress (5)	[703]
COG0533, YgjD (At2g45270, At4g22720)	t^6A biosynthesis/telomere maintenance	In progress (1)	[730]
COG0009, YciO (At3g01920)	RNA metabolism	In progress (1)	[223]
COG0523 (At1g15730, At1g26520, At1g80480)	Metal chaperone, metal homeostasis	In progress (7)	[1135]
COG2016 At1g09150	acp^3 U in RNA	In progress (1)	[40]
COG2263 (At4g28830)	RNA methylase m^6A in rRNA	In progress (1)	[40]
PROSC, YggS (At4g26860, At1g11930)	Proline biosynthesis	Pending	[622]
COG0624 (At5g43600)	Alternative N-formylglutamate deformylase	Pending	[26]
COG0697 (At3g02690)	Pterin efflux carrier	Pending	[36]
COG0451 (At1g19690)	Synthesis of sugars decorating lipid A	Pending	[436]
COG0277 (At4g36400)	Hydroxyacid dehydrogenase	Pending	[176]

Comparative Genomics and Experimental Validation to Find Universal, Globally Missing Genes: The Universal Families COG009 and CO0533

Basma El Yacoubi[1]* (basma@ufl.edu) and **Valérie de Crécy-Lagard**[1] (vcrecy@ufl.edu)

[1]Dept. of Microbiology and Cell Science, University of Florida, Gainesville, Fla.

By combining comparative genomics-guided functional predictions based on plants and prokaryotes with experimental validations, we plan to predict and experimentally verify the function of 20 unknown protein families. This will permit annotation of >3000 individual genes. (See Valérie de Crécy-Lagard and Andrew D. Hanson presentations and poster by Anne Pribat). Here, we illustrate the approach with the case of two universal and globally missing genes. We exploited the availability of whole-genome sequence data, especially those of microorganisms with small genomes, to investigate the minimal gene set, and to find potentially globally missing genes for which biochemical or physiological data suggests universality.

t^6A is a universal base modification occurring at position 37 in a subset of tRNAs decoding ANN codons. The biosynthesis pathway of this complex modification is not elucidated but is known to require threonine, ATP and bicarbonate (1,2). To date no gene has been associated with the t^6A_{37} biosynthesis pathway. Because it is universal and globally missing, traditional boinformatic tools such as blast, phylogenetic occurrence and physical clustering cannot be applied to guide functional predictions. Instead, we generated a list of ortholog families present in a subset of the smallest genomes (total 16 genomes). This analysis generated a list of 95 orthologous families, only nine of these families did not have an experimentally verified function when this work was initiated and we focused on two of them for potential missing t^6A genes for the reasons developed below.

1. Identification of COG0009 as a potential threonylcarbamoyladenosine (t^6A) enzyme

The COG0009 family was chosen as the first candidate for an involvement in t^6A biosynthesis, because it has been shown to bind double-stranded RNA (3) and has been linked to translation phenotypes in both prokaryotes and eukaryotes (4,5), and also because of sequence homology with [NiFe] hydrogenase maturation protein HypF, which catalyzes a reaction with a chemistry similar to the one expected for a t^6A enzyme (6).This gene family can further be split based on sequence comparison into three subfamilies: YrdC, Sua5 (YrdC with an extra domain termed Sua5) and YciO (Figure 2). One or two members of this family are present in each genome. The *A. thaliana* and *E. coli* genomes for example contain two, YrdC and YciO, while the yeast genome contains only one homolog, Sua5. We showed that 1) tRNAs from strains of *S. cerevisiae* lacking a *sua5* do not

contain t^6A and that this phenotype is complemented by transforming with a plasmid encoding the wild type gene; 2) the homologs from *B. subtilis*, *M. maripaludis*, *E. coli yrdC*, but not the *E. coli yciO* are able to complement the t^6A minus phenotype of the yeast *Δsua5* and 3) the *yrdC* homolog is essential in *E. coli*, whereas *yciO* is not, and yeast lacking *sua5* are greatly impaired in their growth (information which was controversial in the literature); 4) *S. cerevisiae*, *B. subtilis*, *M. maripaludis* are able to complement the lethality phenotype of *yrdC* in *E coli*, but not *E. coli yciO* and 5) *E. coli yrdC* is able to bind t^6A apomodified $tRNA^{Thr}$ but not unmodified transcript. Therefore, members of the YrdC/Sua5 family are most probably involved in t^6A biosynthesis. It is unclear at this point which of the homologs of *A. thaliana* is a functional YrdC, and work is underway to investigate the plant family.

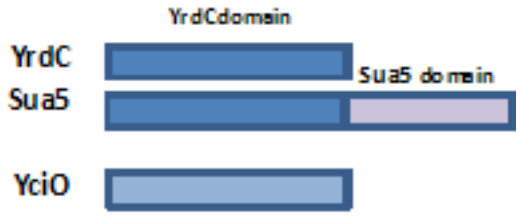


Figure1. Domain organization of the YrdC/Sua5 family. Sua5 members have a well conserved YrdC domain but contain an addition domain, the Sua% domain found only in the family of proteins. The YciO members are missing some conserved residues of the YrdC domain.

2. Identification of the COG0533 as another potential t^6A enzyme

The biosynthesis of t^6A requires multiple enzymatic steps and therefore, partners of YrdC should also be implicated in this pathway. These should follow the same phylogenetic distribution as COG0009. A candidate also identified through our bioinformatic analysis is COG0533. Interestingly, the domain conserved in COG0533 (YgjD domain) is also found in HypF and some HypF proteins contain both a YrdC and a YgjD domain. Also a YgjD domain is found in NodU and NolO which are carbamoyl transferases, chemistry that would resemble one of a t^6A enzyme (Figure 2).

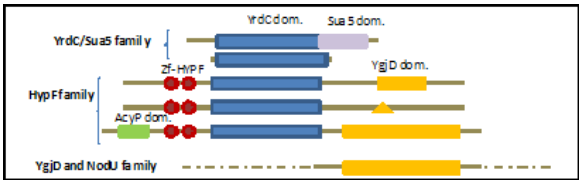


Figure 2. Conserved domain architecture of YrdC domain containing protein relevant to the t6A biosynthesis pathway. This was based on NCBI CDART and (1).
*Possible additional domains not shown.

To validate this family as a t^6A biosynthesis enzyme, t^6A content will be analyzed in yeast and *A. thaliana ygjD* mutants and homologs from appropriately chosen Bacteria and Archaea will be tested for functional complementation. Also i*n vitro* assays will be performed with recombinant YrdC and YgjD to confirm their roles as partners in t^6A biosynthesis. Finally the YgjD family in Eukaryotes contains two sub-families, the bacterial-like YgjD family (yeast YDL104C) and the archaeal and eukaryotic family (yeast YKR038C). The latter has been implicated in telomere length regulation (7). A member of each family from *Arabidopsis thaliana* and yeast will be tested for their potential role in t^6A biosynthesis.

References

1. Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *Journal of Molecular Biology*, 287, 1023-1040.
2. Elkins, B.N. and Keller, E.B. (1974) The enzymatic synthesis of N-(purin-6-ylcarbamoyl)threonine, an anticodon-adjacent base in transfer ribonucleic acid. *Biochemistry*, 13, 4622-4628.
3. Teplova, M., Tereshko, V., Sanishvili, R., Joachimiak, A., Bushueva, T., Anderson, W.F. and Egli, M. (2000) The structure of the *yrdC* gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. *Protein Sci.*, 9, 2557-2566.
4. Kaczanowska, M. and Ryden-Aulin, M. (2005) The YrdC protein: a putative ribosome maturation factor. *Biochim. Biophys. Acta*, 1727, 87-96.
5. Na, J.G., Pinto, I. and Hampsey, M. (1992) Isolation and characterization of *sua5*, a novel gene required for normal growth in *Saccharomyces cerevisiae*. *Genetics*, 131, 791-801.
6. Paschos, A., Bauer, A., Zimmermann, A., Zehelein, E. and Bock, A. (2002) HypF, a carbamoyl phosphate-converting enzyme involved in [NiFe] hydrogenase maturation. 277, 49945-49951.
7. Downey, M., Houlsworth, R., Maringele, L., Rollie, A., Brehme, M., Galicia, S., Guillard, S., Partington, M., Zubko, M.K., Krogan, N.J. *et al.* (2006) A genome-wide screen identifies the evolutionarily conserved KEOPS complex as a telomere regulator. *Cell*, 124, 1155-1168.

GTL

Genome Annotation: Coupling the Power of Plant-Prokaryote Comparative Genomics to Experimental Validation, COG0720 and COG3404

Anne Pribat[1]* (pribat@ufl.edu), Linda Jeanguenin,[1] **Valérie de Crécy-Lagard[2]** (vcrecy@ufl.edu), and Andrew D. Hanson[1] (adha@ufl.edu)

[1]Dept. of Horticultural Sciences and [2]Dept. of Microbiology and Cell Science, University of Florida, Gainesville, Fla.

Project Goals: By combining comparative genomics-guided functional predictions based on plants and prokaryotes with experimental validations, we plan to predict and experimentally verify the function of 20 unknown protein families. This will permit annotation of >3000 individual genes.

The recent rapid expansion of genome sequencing projects has produced a mass of sequence data that are largely annotated automatically. With more than 800 genomes now completed and over 3,000 more in the pipeline, the propagation of unhelpful or incorrect annotations (e.g., 'hypothetical protein', generic annotations, annotation errors) is increasing exponentially. This degrades databases and so prevents optimal use of genome information.

By combining comparative genomics-guided functional predictions based on plants and prokaryotes with experimental validations, we plan to predict and experimentally verify the function of 20 unknown protein families. This will permit annotation of >3000 individual genes.

Here, we illustrate the approach with unknown protein families COG0720 and COG3404.

1. COG0720: A novel enzyme in folate biosynthesis
Certain lower plants (diatoms, pelagophytes) and many bacteria lack the folate synthesis enzyme dihydroneopterin aldolase (FolB), which converts dihydroneopterin to 6-hydroxymethyl-dihydropterin (HMDHP) in the folate biosynthesis pathway (Fig. 1). Comparative genomic analysis showed that most plants and bacteria that lack FolB have a paralog (COG0720) of the tetrahydrobiopterin synthesis enzyme 6-pyruvoyltetrahydropterin synthase (PTPS) in which a glutamate replaces or accompanies the canonical catalytic cysteine residue. A similar COG0720 protein from the malaria parasite *Plasmodium falciparum* (which is related to plants) was recently shown to form HMDHP from dihydroneopterin triphosphate *in vitro*, and was proposed to provide a bypass to the FolB step *in vivo* (Fig. 1) (Dittrich *et al. Mol. Microbiol.* **67:** 609).

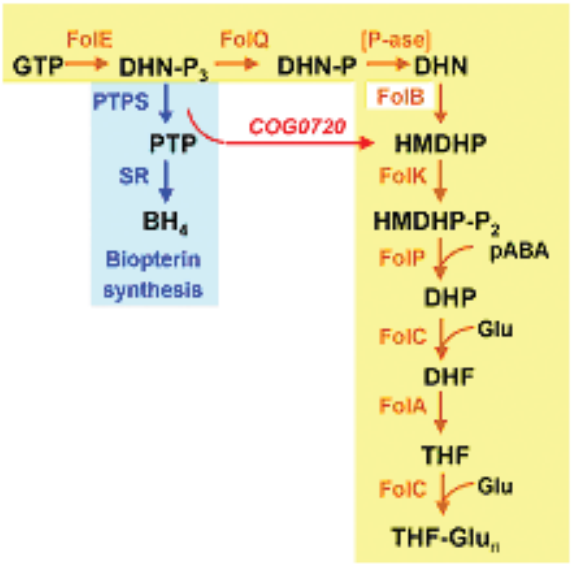


Figure1. Folate synthesis pathway

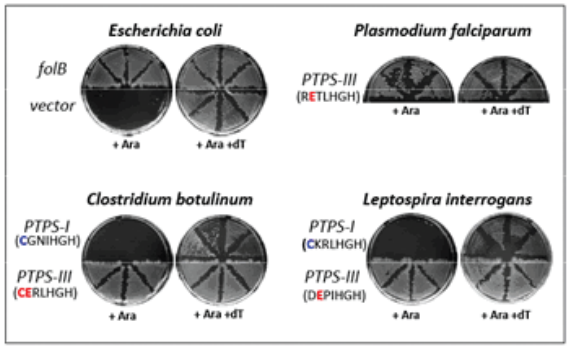


Figure 2. *E. coli* ΔfolB complementation by PTPS-III proteins

Bacterial COG0720 genes encoding proteins with active-site glutamate or glutamate plus cysteine (PTPS-III proteins), or cysteine alone (PTPS-I proteins), were accordingly tested along with the *P. falciparum* sequence for complementation of the *Escherichia coli folB* mutation. The *P. falciparum* sequence and other PTPS-III proteins were active; PTPS-I proteins were not (Fig. 2). These results establish that COG0720 proteins indeed functionally replace FolB.

2. COG3404: A probable novel folate metabolism enzyme
Folates are essential as cofactors for one-carbon (C_1) transfer reactions in nucleotide biosynthesis, amino acid metabolism, and elsewhere. 5-Formyltetrahydrofolate (5-CHO-THF), formed by a side-reaction of serine hydroxymethyltransferase (SHMT), is not a C_1 donor but a potent inhibitor of various folate-dependent enzymes. It must consequently be metabolized. The only enzyme known to do this is 5-CHO-THF cycloligase (5-FCL, Fig. 3). However, there are almost surely alternative enzymes that metabolize 5-CHO-THF because knocking out 5-FCL in plants and microorganisms gives rather mild phenotypes (Goyer *et al.*, *J. Biol. Chem.* 280: 26137; Holmes *et al.*, *J. Biol. Chem.* 277: 20205). Comparative genomics analysis identified a strong candidate: COG3404, a formiminotransferase-like protein, which potentially mediates transfer of the formyl group of 5-CHO-THF to glutamate. This prediction can be validated by testing whether COG3404 substitutes for 5-FCL in complementation assays using an *E. coli* 5-FCL (*ygfA*) deletant.

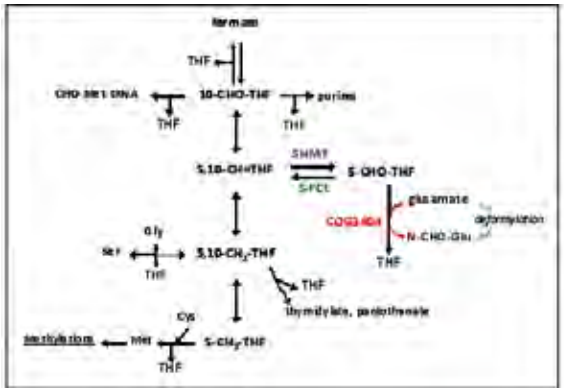


Figure 3. Folate metabolism overview

GTL

Use of Modern Chemical Protein Synthesis Techniques to Experimentally Validate the Functional Annotation of Microbial Genomes

Stephen Kent* (skent@uchicago.edu) and Youhei Sohma

Institute for Biophysical Dynamics, University of Chicago, Chicago, Ill.

Project Goals: To develop high throughput chemical methods to make large numbers of predicted proteins and protein domains, based on microbial genome sequences.

Chemical protein synthesis is a powerful way of studying the properties of predicted proteins. It involves the use of organic chemistry to construct a predicted polypeptide chain from protected amino acid starting materials, followed by folding of the synthetic polypeptide to give the unique, defined tertiary structure of the protein molecule. The synthetic protein is then used to experimentally validate the predicted biochemical function, and in selected cases to determine the Xray structure of the protein molecule (**Figure 1**).

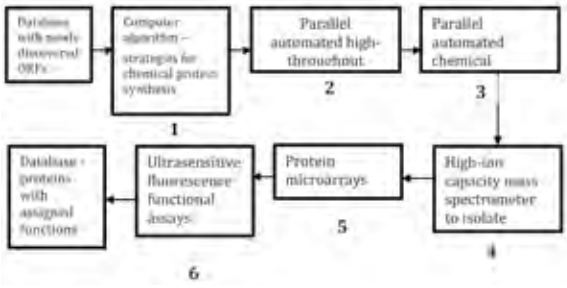
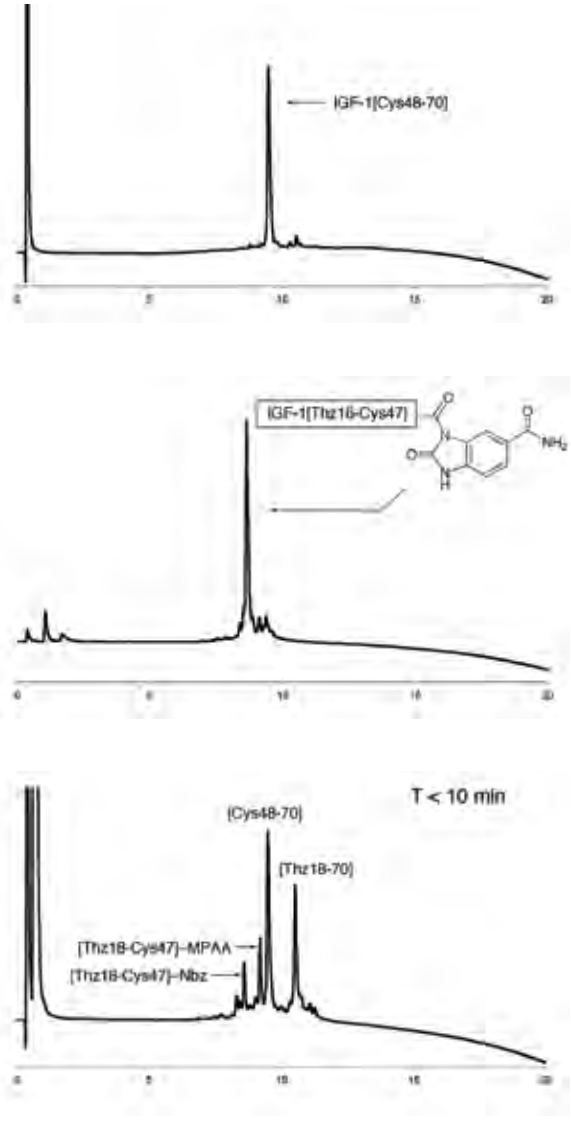


Figure 1. Modular high-throughput platform for fast and parallel total chemical synthesis, mass-spectrometric purification and single-molecule spectroscopic assay to annotate function for newly predicted proteins.

Modern total protein synthesis has evolved from the 'chemical ligation' methods introduced by the Kent laboratory in the mid-1990s [Dawson PE, Kent SB: Synthesis of native proteins by chemical ligation. *Annual Review of Biochemistry* 2000, **69**:923-960.]. Unprotected synthetic peptide segments, spanning the amino acid sequence of the target polypeptide chain, are covalently joined to one another by chemo-selective reaction of unique, mutually reactive functional groups on each segment. Native chemical ligation, the thioester-mediated covalent bond-forming chemoselective reaction of unprotected peptides at a Cys residue, is the most robust and useful ligation chemistry developed to date.

Synthesis of peptide-thioesters. For many biomedical researchers the utility of chemical protein synthesis based on native chemical ligation methods is limited by the inability to make peptide-thioesters. Routine synthesis of peptide-thioesters has not previously been possible using Fmoc chemistry SPPS methods [Camarero JA, Mitchell AR: Synthesis of proteins by native chemical ligation using Fmoc-based chemistry. *Protein and Peptide Letters* 2005, **12**:723-728]. In the first stage of prototyping high throughput chemical protein synthesis, we have used x,y,z robotics and laboratory automation to develop efficient Fmoc chemistry SPPS protocols for the simultaneous parallel synthesis of the key peptide-thioester building blocks needed for chemical protein synthesis. This made use of a recently reported novel resin linker [Blanco-Canosa JB, Dawson PE: An efficient Fmoc-SPPS approach for the generation of thioester peptide precursors for use in native chemical ligation. *Angew Chem Int Ed Engl.* 2008, **47**:6851-5]. Typical data are shown in **Figure 2 (Top)**.



* Presenting author

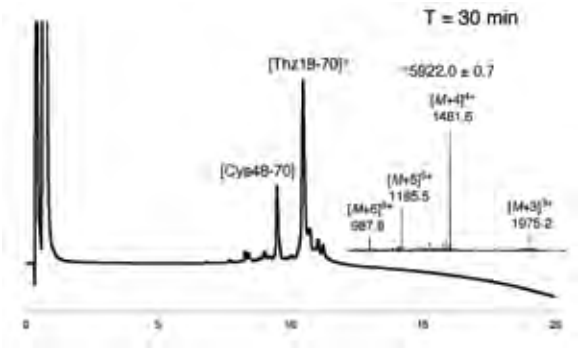


Figure 2. (Top): automated robotic Fmoc SPPS preparation of peptide-thioesters. HPLC-electrospray MS of crude products are shown. (Bottom): LCMS analyses of native chemical ligation of the two peptides; near-quantitative reaction after 30 minutes was observed. The 53 amino acid residue ligated product had a mass of 5922 Daltons.

We expect that this approach to parallel peptide synthesis will be satisfactory for the simultaneous production of ~100 peptide-thioester segments, each containing 30 to 40 amino acid residues. Ready preparation of peptide-thioesters enables the straightforward total chemical synthesis of proteins by native chemical ligation **Figure 2 (Bottom)**. Proof-of-concept total chemical syntheses of several predicted proteins from microbial genomes will be presented.

GTL

Towards Annotation of the Unannotated—A Dissection of Unannotated Proteins in *Clostridium thermocellum*

Y.J. Chang,* Loren J. Hauser, Gwo-liang Chen, Frank Larimer, Pavan Ghatty, and **Miriam Land** (ml3@ornl.gov)

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

Project Goals: In all prokaryotic genomes sequenced there is a substantial fraction of genes that are uncharacterized or unknown, which poses one of the biggest hurdles to the development of genome annotation, thus genome science. The annotation of unknown proteins remains an unsolved problem.

Attempts to solve this problem have been limited, due to the lack of methodology, also the overwhelming interest in immediate opportunity of analyzing known sequences. In this study, we systematically explored the unannotated genes in genome of *Clostridium thermocellum* ATCC 27405. The approaches we applied include genomic context, gene order comparison, protein structure and phylogenetic analysis. With the help of extensive searches across all other genomes and horizontal gene transfer (HGT)/transposable elements (TE) /repeat detection, clusters of functionally related unknown genes are identified. A possible relationship between unknown and other neighboring genes, such as HGT, protein splicing site, transposable elements is discussed.

This whole genome initiative aimed at 1) initiating annotation of the unannotated at genome scale 2) mapping out the function and possibly regulatory relationships of unknown proteins, and 3) eventually defining the uncharacterized proteins. The work presented here represents a part of the ongoing efforts to annotate the unannotated at the Oak Ridge National Laboratory.

In all prokaryotic genomes sequenced there is a substantial fraction of genes (up to 60%) that are uncharacterized or unknown, which poses one of the biggest hurdles to the development of genome annotation, thus genome science. The annotation of unknown proteins remains an unsolved problem.

Attempts to solve this problem have been limited, due to the lack of methodology, the overwhelming interests in immediate opportunity of analyzing known proteins. In this study, we systematically explored the unannotated genes in genome of *Clostridium thermocellum* ATCC 27405. The approaches applied include genomic context, gene order comparison, protein structure and phylogenetic analysis. With the help of extensive searches across all other genomes and horizontal gene transfer (HGT)/repeat detection, clusters of functionally related unknown genes are identified. A possible relationship between unknown and other neighboring genes, such as HGT, protein splicing site, transposable elements is discussed.

This whole genome initiative aimed at 1) initiating annotation of the unannotated at genome scale 2) mapping out the function and possibly regulatory relationships of unknown proteins, and 3) eventually defining the uncharacterized proteins. The work presented here represents a part of the ongoing efforts to annotate the unannotated at the Oak Ridge National Laboratory.

GTL

An Integrated Approach to Experimental Validation of Putative Gene Functions in *M. acetivorans*

Ethel Apolinario,[1] Yihong Chen,[2] Zvi Kelman,[2] Zhuo Li,[2] Basil J. Nikolau,[3] Kevin Sowers,[1] Eric Testroet,[3] and **John Orban**[2,*] (orban@umbi.umd.edu)

[1]Center of Marine Biotechnology and [2]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Md.; and [3]W. M. Keck Metabolomics Research Laboratory, Iowa State University, Ames, Iowa

Project Goals: The goal of the project is to develop an integrated high-throughput approach to functionally

annotate a large group of conserved hypothetical genes in the methanogenic archaeon, *Methanosarcina acetivorans*. The focus will be on genes predicted to encode enzymes, the substrate(s) and products of which are unknown. Approximately 2226 of the 4524 genes in *M. acetivorans* fall into this category and include genes possibly involved in processes such as methanogenesis, nitrogen fixation, and carbon assimilation. The biochemical functions of these putative enzymes will be accurately annotated using a combination of gene knockouts, high throughput metabolomic analysis with mass spectrometry (MS), automated screening of implicated metabolites with nuclear magnetic resonance spectroscopy (NMR), and biochemical assays.

We are developing an integrated approach to facilitate experimental determination or validation of enzymatic functions in the methanogenic archaeon, *Methanosarcina acetivorans* (MA). The goal is to use a combination of gene knockouts, metabolic profiling with mass spectrometry, NMR-based ligand screening, and biochemical assays to accurately annotate gene targets.

Before choosing suitable initial targets for experimental validation, we first manually re-annotated ~710 of the 4524 MA genes based on updated annotations in other Archaea and a thorough search for experimental data in the literature. This process led to ~220 gene annotations which were more specific than the original function assignment, ~60 gene annotations became less specific, and ~430 gene annotations were not altered. The set of genes with a more specific functional annotation was further surveyed for 1) putative enzymatic function, 2) *in vivo* expression in MA from existing proteomic and DNA microarray data, and 3) transmembrane containing regions. Approximately 30 of these gene products were putatively identified as soluble enzymes with detectable *in vivo* expression and were targeted for cloning, expression screening and purification. Genes that could be expressed at reasonable levels of soluble protein in *E. coli* were further targeted for preparation of knockout alleles in MA. Alternatively, if an *E. coli* homolog exists, a non-essential gene knockout will be obtained directly from the Keio collection.

Expressed proteins are screened for putative substrates and products using a one-dimensional ^{1}H NMR-based assay, waterLOGSY (water-Ligand Observed via Gradient Spectroscopy), which has been used in the past for drug screening applications (*1*). Additionally, we are using mass spectrometry to profile metabolite differences in *E. coli* strains containing the relevant MA gene and we will profile metabolites in MA gene knockout strains. Genetic complementation of *E. coli* knockout strains is also being used to confirm the function of MA genes. Results on genes putatively identified as being involved in biotin metabolism and regulation will be highlighted.

Reference

1. Dalvit, C., Fogliatto, G., Stewart, A., Veronesi, M., and Stockman, B. (2001) WaterLOGSY as a method for primary NMR screening: practical aspects and range of applicability, *J Biomol NMR* 21, 349-359.

GTL

Transcriptome Analysis of *Chlamydomonas reinhardtii* using Ultra-High-Throughput Sequencing

David Casero Díaz-Cano,[1] Madeli Castruita,[2] Sabeeha Merchant,[2] and **Matteo Pellegrini**[1]* (matteop@mcdb.ucla.edu)

[1]Dept. of Molecular, Cell and Developmental Biology and [2]Dept. of Chemistry and Biochemistry, University of California, Los Angeles, Calif.

Project Goals: The advent of massively parallel short read sequencing technology opens the door to (near) full coverage of the *Chlamydomonas* transcript map via deep sequencing of mRNAs. To evaluate the potential of Illumina's Solexa technology for a) generating a whole transcriptome for *Chlamydomonas*, b) identifying differentially expressed genes, and c) reconstructing gene models de novo, we analyzed RNAs isolated from metal deficient conditions and developed novel algorithms for data analysis.

Chlamydomonas reinhardtii, a unicellular eukaryote in the plant lineage, has been exploited in the laboratory over the last 50 years as a model organism for the study of eukaryotic photosynthesis. Unlike flowering plants, *Chlamydomonas* synthesizes and maintains a functional photosynthetic apparatus even when grown in the dark by respiration on organic carbon. This means that mutants with defects in either the light or dark reactions of photosynthetic metabolism can be maintained and characterized biochemically. Accordingly, many of the fundamental discoveries leading to today's knowledge of photosynthesis are derived from the application of biochemical and classical genetic approaches using the *Chlamydomonas* model. In the last decade, pathways of energy metabolism beyond photosynthesis have received considerable attention from the research community, specifically the biosynthesis of H_2 and the production of ethanol and other fermentation products resulting ultimately from solar energy conversion in the photosystems.

The genome of *Chlamydomonas* consists of 121 Mb in 17 chromosomes. Relative to other eukaryotes, a typical *Chlamydomonas* gene is intron-rich; there are 8.3 exons per gene and the average intron size is 373 bp. These characteristics make de novo prediction of gene models very difficult in the absence of a high quality dense transcript map. The existing datasets cover only 8631 (about half) of the 15,143 predicted protein-coding gene models, and only half of these include full-length coverage. Accordingly, despite the importance of *Chlamydomonas* as a model for the study of photosynthesis and energy metabolism, **only a quarter of the protein-coding gene models are accurately computed and verified via a transcript map.**

The advent of massively parallel short read sequencing technology opens the door to (near) full coverage of the *Chlamydomonas* transcript map via deep sequencing of mRNAs. To

* Presenting author

evaluate the potential of **Illumina's Solexa technology** for a) generating a whole transcriptome for *Chlamydomonas*, b) identifying differentially expressed genes, and c) reconstructing gene models *de novo*, we analyzed RNAs isolated from metal deficient conditions

We have verified that these libraries may be used to quantitatively estimate transcript fold changes in different conditions using existing gene models. We are also developing a new annotation pipeline using only the short read sequencing data, and have shown that even simple approaches allow us to accurately reconstruct a set of manually curated genes. This approach therefore promises to not only measure transcript counts and differential expression but also comprehensively annotate the genomes of organisms for which we have only partial genome sequences.

GTL

Transcript Verification Coupled with Metabolic Network Modeling for *Chlamydomonas reinhardtii*

Ani Manichaikul,[1]* Lila Ghamsari,[2,3] Chenwei Lin,[2,3] Erik F.Y. Hom,[4] Ryan R. Murray,[2,3] Roger L. Chang,[5] Arvind K. Chavali,[1] Yun Shen,[2,3] Xinping Yang,[2,3] Ines Thiele,[5] Jason A. Papin[1] (papin@virginia.edu), and **Kourosh Salehi-Ashtiani**[2,3] (Kourosh_Salehi-ashtiani@DFCI.harvard.edu)

[1]Dept. of Biomedical Engineering, University of Virginia Health System, Charlottesville, Va.; [2]Center for Cancer Systems Biology (CCSB), Dept. of Cancer Biology, Dana-Farber Cancer Institute, Boston, Mass.; [3]Dept. of Genetics, Harvard Medical School, Boston, Mass.; [4]Dept. of Molecular and Cellular Biology, Harvard University, Cambridge, Mass.; [5]Dept. of Bioengineering, University of California San Diego, La Jolla, Calif.

Project Goals: Our objectives for this project are to experimentally verify, define, and validate metabolic protein-coding genes of *Chlamydomonas reinhardtii* and model a comprehensive metabolic network for this organism. The proposed experiments utilize a technology platform that can be adapted to virtually any organism, and hence serve as a prototype that can be used for gene validation in any species. *Chlamydomonas reinhardtii* is an ideal organism for this project because 1) it is an important "bio-energy" organism, and 2) a draft of its genome sequence is currently available. The obtained results will be used to build a more complete model of the metabolic circuitry of this organism. The generation of a metabolic network will in turn help validate examined genes by defining a biological role for them. From our obtained results, we should be able to formulate testable hypotheses as to how to optimize bio-fuel (including hydrogen gas) production in this organism. To achieve these objectives, we will be carrying out experiments to define and verify transcript structures of metabolic genes in *Chlamydomonas reinhardtii* by RT-PCR and RACE, functionally validate the transcripts by yeast two-hybrid experiments, and build and interpret predictive metabolic network models based on the obtained results.

With genome sequencing of many bioenergy organisms completed or in progress, there is a growing need to bridge the gaps between primary sequence information, gene annotation, and mapping of metabolic networks. The release of *C. reinhardtii* genome sequence has made this unicellular algae a viable target for metabolic engineering towards improved biofuel production; however, the relevant genes are not validated and the needed metabolic network map is not currently available. We present an update on our integrative process of coupling network reconstruction with transcript verification on *C. reinhardtii*. Using literature (and other resources) we have reconstructed a model of the central metabolic network of *C. reinhardtii*. Our *in silico* reconstructed network encompasses 252 reactions, 109 metabolites, and 295 gene products, and accounts for their intracellular compartmentalizations. Through systematic comparison of *JGI* transcript annotation against publicly available protein sequence databases, we identified and assigned E.C. numbers to transcript sequences of all but 12 of the hypothesized gene-associated reactions. To validate the involved gene products, we have carried out open reading frame (ORF) verification by RT-PCR and RACE (rapid amplification of cDNA ends) for all the protein coding genes involved (as well as a set of positive control ORFs). Following optimization of the RT-PCR procedure for high GC content of *C. reinhardtii* transcriptome, we were able to verify (by cloning and sequencing) approximately 80% of the ORFs annotated in the central metabolic map. Data from RACE experiments is being used to further verify and refine annotation of the transcripts. The generated metabolic model, now supported by experimental results, carries *in silico* 'growth' with a yield of 0.012 g DW / mmol acetate, consistent with the experimentally derived value for the organism. The network, along with ORF verification experiments is now being expanded to include all major metabolic reactions, including (but not limited to), fatty acid, isoprenoid, carotenoid and other hydrocarbon pathways essential to certain biofuel production. The metabolic ORF clone resource that we have generated will be made available without restrictions to the research community.

Genemap-MS: High Throughput Mass Spectrometry Approaches to Microbial Gene Annotation Validation

Trent R. Northen* (trnorthen@lbl.gov), Benjamin Bowen, Steven M. Yannone, Jill Fuss, John Tainer, and **Gary Siuzdak** (siuzdak@scripps.edu)

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Project Goals: Here an integrated approach is used to develop universally applicable high-throughput (HT) methods for validating genome annotation using mass spectrometry (MS) based proteomics, metabolomics, and our developing technologies for detecting biochemical activities on arrayed metabolite substrates. This leverages our expertise with developing and applying MS technologies to generate datasets directly applicable to validating currently annotated genomes. The technologies represent multifaceted and universally applicable approaches to validate: 1) protein expression 2) metabolic pathways and 3) biochemical activities. The datasets from these analyses are integrated into computational metabolic networks to provide functional validation of the many hypothetical activities in current genome annotations. We provide a balance of mature and robust MS technologies with new surface based MS technologies and the expansion of our METLIN metabolite database with the ultimate goal of addressing specific DOE needs for exploiting microbes for bioenergy production.

The utility of genetic information being derived from sequencing efforts is diminished by the incomplete and sometimes incorrect annotations associated with "completed" genomes. Homology-based protein function predictions are limited by evolutionary processes that result in conserved domains and sequence being shared by enzymes of widely diverse functions. Therefore, additional experimental datasets directed at validating and improving genome annotations are required. Project Genemap-MS is developing and applying universally applicable high-throughput (HT) methods for validating genome annotation using mass spectrometry (MS) based proteomics, metabolomics, and our developing technologies for detecting biochemical activities on arrayed metabolite substrates (NIMS and Nimzyme). To maximally improve existing homology based annotations, we are using diverse model systems which span the three branches of the tree of life *Synechococcus sp* (cyanobacteria), *Sulfolobus solfataricus* (archaea), *Chlamydomonas Reinhardtii* (eukaryota); an integrated metabolomics/proteomics pipeline; and our new Meta-IQ bioinformatics software.

Novel Mass Spectrometry Based Platforms for the Investigation of Model Organisms

Sunia A. Trauger[1]* (strauger@scripps.edu), Trent Northen,[2] Nitin Baliga,[3] Steven M. Yannone,[2] Michael W.W. Adams,[3] and **Gary Siuzdak**[1]

[1]Dept. of Molecular Biology and Center for Mass Spectrometry, The Scripps Research Institute, La Jolla, Calif.; [2]Dept. of Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [3]Institute for Systems Biology; University of Georgia, Athens, Ga.

Project Goals: The objective of the MAGGIE Program is to comprehensively characterize the Protein Complexes (PCs) and Modified Proteins (MPs) underlying microbial cell biology. MAGGIE will address immediate GTL missions by accomplishing three specific goals: (1) to provide a comprehensive, hierarchical map of prototypical microbial PCs and MPs by combining native biomass and tagged protein characterizations from hyperthermophiles with comprehensive systems biology characterizations of a non-thermophilic model organism, (2) to develop and apply advanced mass spectroscopy and SAXS technologies for high-throughput characterizations of PCs and MPs, (3) to create and test powerful computational descriptions for protein functional interactions.

The Scripps Center for Mass Spectrometry is focused on developing methodologies for the characterization of protein-protein complexes and the elucidation of functional pathways affected by specific perturbations for the three model organisms of interest: ***Pyrococcus furisiosus***, ***Sulfolobus solfataricus*** and ***Halobacterium salinarum***. The different mass spectrometry based platforms we have developed include, capillary LC based metabolomics, high-throughput proteomics approaches as well as nanostructure initiator mass spectrometry (NIMS) which is a new mass spectrometry based approach which shows considerable promise for the monitoring of a specific enzymatic activity in microbial communities.

The field of global mass-based metabolomics provides a platform for discovering unknown metabolites and their specific biochemical pathways. We report the identification of a new endogenous metabolite, N(4)-(N-acetylaminopropyl)spermidine which was found to be up-regulated and the use of a novel proteomics based method for the investigation of its protein interaction using metabolite immobilization on agarose beads. The metabolite was isolated from the organism *Pyrococcus furiosus*, and structurally characterized through an iterative process of synthesizing candidate molecules and comparative analysis using accurate mass LC-MS/MS. An approach developed for the selective preparation of N(1)-acetylthermospermine, one of the possible structures of the unknown metabolite, provides a convenient route to new polyamine derivatives through methylation on the

N(8) and N(4) of the thermospermine scaffold. The biochemical role of the novel metabolite as well as that of two other polyamines: spermidine and agmatine is investigated through metabolite immobilization and incubation with native proteins. The identification of eleven proteins that uniquely bind with N(4)-(N-acetylaminopropyl)spermidine, provides information on the role of this novel metabolite in the native organism. Identified proteins included hypothetical ones such as PF0607 and PF1199, and those involved in translation, DNA synthesis and the urea cycle like translation initiation factor IF-2, 50S ribosomal protein L14e, DNA-directed RNA polymerase, and ornithine carbamoyltransferase. The immobilization approach demonstrated here has the potential for application to other newly discovered endogenous metabolites found through untargeted metabolomics, as a preliminary screen for generating a list of proteins that could be further investigated for specific activity.

We have also developed a Nanostructure-Initiator Mass Spectrometry (NIMS) enzymatic (Nimzyme) assay in which enzyme substrates are immobilized on the mass spectrometry surface by using fluorous-phase interactions. This "soft" immobilization allows efficient desorption/ionization while also enabling the use of surface-washing steps to reduce signal suppression from complex biological samples, which results from the preferential retention of the tagged products and reactants. The Nimzyme assay is sensitive to subpicogram levels of enzyme, detects both addition and cleavage reactions (sialyltransferase and galactosidase), is applicable over a wide range of pHs and temperatures, and can measure activity directly from crude cell lysates. The ability of the Nimzyme assay to analyze complex mixtures is illustrated by identifying and directly characterizing β-1,4-galactosidase activity from a thermophilic microbial community lysate.

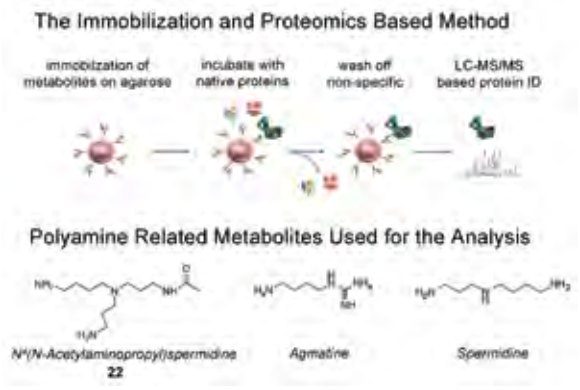


Figure 1. The metabolite immobilization, incubation with cell lysate, followed by proteomic analysis allows a first look at identifying proteins which interact with a novel metabolite of interest.

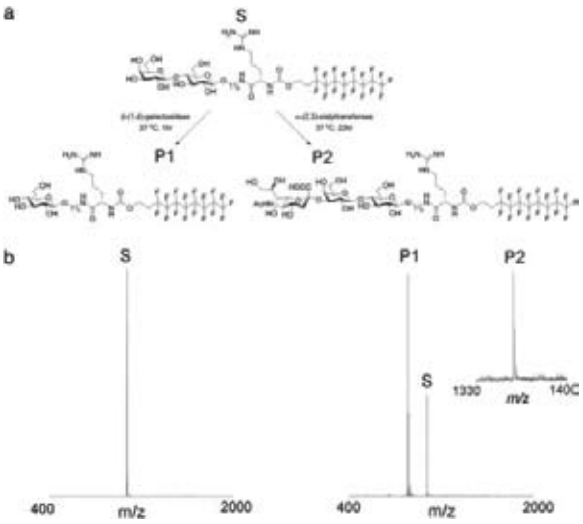


Figure 2. On-chip NIMS enzymatic activity assay (Nimzyme assay). (*a*) Substrate (*b*) Mass spectra of the substrate (*Left*) and resulting products (*Right*).

GTL

Annotation of Translation Initiation Sites Using Prodigal

Doug Hyatt, **Miriam Land** (ml3@ornl.gov), and Loren Hauser*

Oak Ridge National Laboratory, Oak Ridge, Tenn.

http://prodigal.ornl.gov/

Project Goals: Provide an update on the microbial gene finding effort at Oak Ridge.

Last year, ORNL introduced the microbial genefinding program Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm). Since that time, Prodigal has been incorporated into the Joint Genome Institute annotation pipeline. Prodigal has been used to annotate all microbial organisms submitted to Genbank by JGI in 2008, resulting in an enormous amount of data with which to measure the accuracy of the algorithm and to make necessary improvements. Prodigal is consistently being improved as new discoveries are made and more data is collected, and new versions are released every couple of months containing the updated changes to the program.

In the course of reviewing the pipeline annotations, we discovered that while Prodigal did quite well on locating translation initiation sites in 85-90% of the genomes, it experienced difficulties in those genomes that do not use the canonical Shine-Dalgarno ribosomal binding site motif (AGGAGG), or some other closely related motif. An effort was made to identify and classify the genomes which do not use this motif (or at least do not use it often). We consulted the literature on translation initiation sites, as well as exam-

ining the full set of finished microbial genomes in Genbank computationally to look for novel motifs. A new version of Prodigal was constructed with a much more complex RBS motif-finding system able to discover novel motifs while not abandoning its knowledge of the default Shine-Dalgarno motif (as happens in many other genefinding programs which auto-discover motifs). This version has since been incorporated into the JGI annotation pipeline.

In consulting the literature, we found that, in *Crenarchaea*, the first gene in an operon often has no ribosomal binding site motif, but genes internal to an operon often do use an SD motif. In *Aeropyrum pernix*, and some other archaea, a GGTG motif was observed computationally. This motif was strong and present in well over 50% of the start sites. Many chlorobi and cyanobacteria were observed to use the SD motif extremely infrequently, and minor AAAA/TATA type motifs were often found 13-15bp upstream (which may be transcriptional in nature). In one organism in particular, the bioenergy-related *Flavobacterium johnsoniae,* literature had documented a strong TAAA motif close to the start codon. This finding was confirmed by our computational analysis. The challenge, after identifying these patterns, was to create a flexible translation initiation site evaluator capable of auto-discovering these novel motifs while not losing sight of the still occasionally used Shine-Dalgarno motif (which could be used so infrequently that it would never be found with motif finding, but is still present in a significant percentage of genes, such as 2-3%).

We approached the problem by creating a motif finder that auto-discovered motifs in the RBS region of length 3-6bp, with the restriction that all 3bp subsets of that motif had to be present in at least 20% of the genes. However, we did allow one mismatch in 5bp and 6bp motifs. The program used an iterative algorithm similar to Prodigal's default Shine-Dalgarno algorithm to assign log-likelihood weights to each motif. We then took this motif finder and Prodigal's default Shine-Dalgarno motif finder and combined them into a single TIS scorer with three distinct cases. In the first instance, if the organism used the SD motif frequently, we used Prodigal's existing default SD scorer. In the second instance, if a novel strong motif, such as GGTG in *Aeropyrum pernix*, was discovered, we used the new scoring system. Finally, if no strong motif of any kind was found, but some weak motifs were found, we used both scoring systems and took the maximum result. In *Crenarchaea*, for example, non-SD motif genes at the beginning of operons will get a decent score from the new scoring system, whereas the internal SD-motif-using genes in operons will get a good score from the old scoring system. This new version of Prodigal was tested on *Cyanobacteria*, *Chlorobi*, *Crenarchaea*, and GGTG-using *Euryarchaea*, and found to outperform the previous version of Prodigal and a version created that only used the new scoring system (but not the old one). The final version of the new TIS finder was completed, and the new version of Prodigal was introduced into the JGI pipeline in December, 2008.

Prodigal is routinely run on all finished genomes in Genbank every couple of months, and detailed comparisons with the Genbank files are performed. This data is available at the Prodigal website (http://prodigal.ornl.gov/).

Computing Resources and Databases

GTL

Release of Taxomatic and Refinement of the SOSCC Algorithm

J. Fish[1]* (fishjord@msu.edu), Q. Wang,[1] S.H. Harrison,[1] T.G. Lilburn,[2] P.R. Saxman,[3] J.R. Cole,[1] and **G.M. Garrity**[1]

[1]Michigan State University, East Lansing, Mich.; [2]American Type Culture Collection, Manassas, Va.; and [3]University of Michigan, Ann Arbor, Mich.

Project Goals: The Taxomatic was conceived as a tool for aiding in the interpretation of large-scale phylogenetic trees, and uncovering unresolved nomenclatural and placement errors in large sets of sequence data using visualization techniques drawn from the field of exploratory data analysis. In this phase of the work we have taken what we have learned from early prototypes and deployed the tool as a web based service available through the RDP.

The Taxomatic was conceived as a tool for aiding in the interpretation of large-scale phylogenetic trees, and uncovering unresolved nomenclatural and placement errors in large sets of sequence data using visualization techniques drawn from the field of exploratory data analysis. The tool allows a user to generate a heatmap image (see Figure 1) representing the pair-wise similarity between large numbers of rRNA sequences (or any other set of homologous genes or concatenated genes) in the context of an existing taxonomy. The Self-Organizing Self-Correcting Classifier (SOSCC) is an algorithm that was developed as an extension to the Taxomatic to programmatically reorder similarity matrices and, based on the reordered matrix, identify and resolve inconsistencies within a proposed taxonomy. Together, these two tools allow a user to: 1) rapidly identify taxonomic anomalies (misplaced sequences), and 2) correct these anomalies.

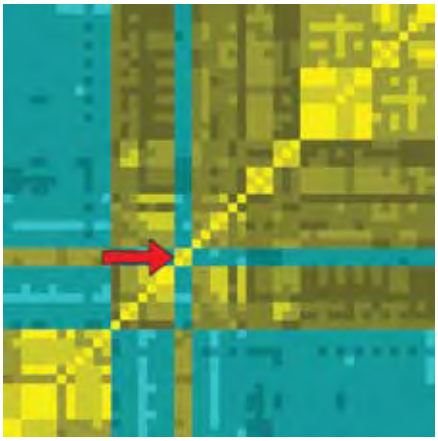


Figure 1. Heatmap of the current taxonomy for the families *Rhodocyclaceae Betaproteobacteria*) and *Brucellaceae* (*Alphaproteobacteria*). A taxonomic anomaly (the genus *Shinella*) can be seen near the center of the image identifiable by the lines of cyan inside the otherwise yellow *Rhodocyclaceae* group.

For input data, the Taxomatic (http://rdp.cme.msu.edu) can use internally retrieved sets of RDP or *my*RDP sequences, externally generated FASTA sequence alignment files, or DNADist files. Within the Taxomatic interface, mouse-over action by a user will trigger the graphical outlining of a taxonomic overlay to highlight specific groups in the input taxonomic hierarchy. The taxonomic hierarchy can either be provided to the user as the default RDP taxonomy or can be directly supplied by the user in XML format. As a third alternative, for the purpose of an unlabelled overview, users may choose to proceed without a selected taxonomy. For higher zoom levels on the heatmap of the Taxomatic interface, detailed strain information is made available to the user through pop-up windows. Users can save the full set of input conditions necessary to either reproduce or extend their current heatmap analysis including: a copy of their heatmap (image size depends on the current zoom level), the matrix used to generate the heatmap, and the taxonomic hierarchy. The taxonomic hierarchy is used as both an algorithmic aid for ordering the sequences in the heatmap, and as a means for generating coordinates on the interface for plotting group overlays.

The SOSCC implements a two-step process in which the source matrix is first reorganized, after which the reordered sequences are reclassified. The source matrix is reorganized in an iterative process that results in an output matrix where closely related sequences are placed adjacent to each other, based on sequence similarity rather than presumptive taxonomic identity. The second phase can proceed along two different approaches. The first approach adjusts the borders of the source taxonomy on the output matrix based on how many adjacent sequences were originally members of the same group. The second approach identifies groups in the output matrix in an unsupervised fashion by identifying large changes in sequence distances. The SOSCC can accept any source data that is acceptable to the Taxomatic and, by default, displays the resulting matrix through the Taxomatic user interface. One option allows for running the SOSCC on 100 bootstrapped samplings drawn from the set of aligned input sequences. With the bootstrapped sampling option, only those taxonomic rearrangements supported by a user-defined threshold level of percentage consistency with the rearrangement will be retained. Bootstrapping is currently limited to 2000 sequences, and can only be used when both the sequence data and taxonomic hierarchy are supplied. When using the bootstrapping feature of the Taxomatic, results are provided to the user via e-mail. These results include: a detailed report on the classification of each sequence; links to view the original and resulting matrix on Taxomatic; and a link to download the resulting matrix.

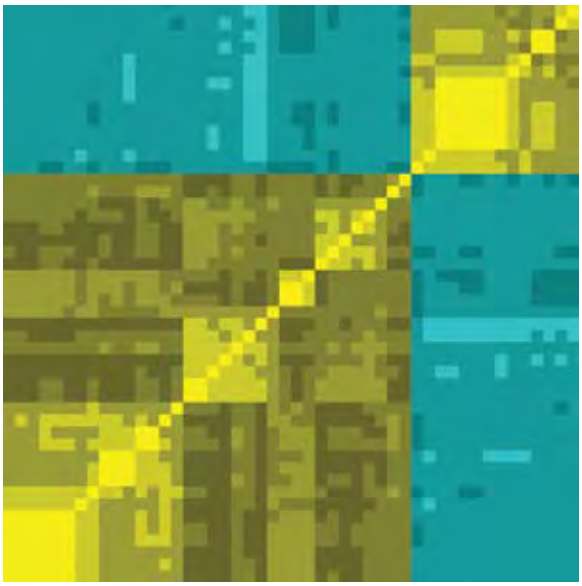


Figure 2. Distance matrix from Fig. 1 preprocessed with SOSCC. *Rhodocyclaceae* (lower left) and *Brucellaceae* (upper right). The discontinuity in the image is gone as *Shinella* has moved to the upper right corner of the image.

The Taxomatic and the SOSCC are both implemented as java servlets, exposing the functionality of their core classes to users using the Spring 2.5 MVC framework. The heatmap image that is displayed by the Taxomatic is delivered as a collection of tiles generated on the fly by the server. The taxonomic hierarchy information is loaded from the server for the currently displayed heatmap. Sequence information is fetched from the NamesforLife (http://names4life.com) database as the user mouse-overs individual sequences at ≥8x zoom levels. All of the presentation code is written using a combination of JSTL and javascript.

Both the Taxomatic and SOSCC have SOAP Web Services ports written in java using the JAX-WS reference implementation. The Taxomatic web service exposes methods to submit a similarity matrix or aligned sequence data to be displayed on the Taxomatic. When aligned sequence data are supplied, an uncorrected distance matrix is calculated from the sequences. The SOSCC web service exposes methods that submit aligned sequence data or a similarity matrix for processing based on either a default set of options or caller-specified set of options.

References

1. J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* (Advance Access November 2008). DOI 10.1093/nar/gkn879.
2. G. M. Garrity, and T, G. Lilburn. Self-organizing and self-correcting classifications of biological data. *Bioinformatics* (2005) 21:2309–2314.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-FG02-04ER63933

SBIR

NamesforLife Semantic Resolution Services for the Life Sciences

Charles T. Parker,[1] Sarah Wigley,[1] Nicole Osier,[2] Jordan Fish,[1] Qiong Wang,[1] Donna McGarrel,[1] James R. Cole,[1,2] Catherine Lyons,[1] and **George M. Garrity**[1,2*] (garrity@names4life.com)

NamesforLife, LLC, and Michigan State University, East Lansing, Mich.

Project Goals: NamesforLife is a novel technology that resolves uncertainty about the meaning of biological names or other dynamic terminologies. It uses those terms to create persistent links to related information, goods, and services available on the Internet, even if the terms have changed.

Within the Genomes-to-Life Roadmap, the DOE recognizes that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpretation of prior data and published results decreases because both become overloaded with synonymous (multiple terms for a single concept) and polysemous terms (single terms with multiple meanings). Ambiguity in rapidly evolving terminology is a common and chronic problem in science and technology.

NamesforLife (N4L) is a novel technology designed to solve this problem. The core consists of an ontology, an XML schema, and an expertly managed vocabulary coupled with Digital Object Identifiers (DOIs) to form a semantic resolution service that disambiguates terminologies, presents them to end-users in a temporal context and persistently links them to relevant resources and services on the Internet. Our initial implementation of N4L technology is for the validly published biological names of *Bacteria* and *Archaea*. These names play a significant role in science, medicine, and government, carry specific meanings to end users in each of those communities, and can trigger responses that may or may not be appropriate in a given situation. Biological names also serve as key terms used to index and access information in databases and the scientific, technical, medical, and regulatory literature. Understanding the correct meaning of a biological name, in the appropriate context, is essential. This is not a trivial task, and the number of individuals with expertise in biological nomenclature is limited. Such knowledge can, however, be accurately modeled and delivered through a networked semantic resolution service that can provide end-users of biological nomenclatures or other dynamic terminologies with the appropriate information, in proper context, on demand. Such a service can also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them more readily discoverable, even when the definition of a name or term has changed.

In our initial studies, we built a prototype to demonstrate that names, concepts, and the objects to which names apply must be treated independently in order to accurately interpret the correct meaning over time. As proof of principle, a preliminary data model and XML schema were developed and a simple semantic resolver was deployed that operated on an underlying dataset of bacterial and archaeal names. In Phase I, that data model was substantially refined to address limitations of the initial prototype and the underlying nomenclatural data were extensively cleaned, extended, and ported to a stable database environment to support development of commercial applications and services. In collaboration with the International Committee on the Systematics of Prokaryotes and the editorial office of the Society for General Microbiology, we built a prototype application of N4L to demonstrate feasibility of high-throughput semantic enablement of scientific literature at the pre-publication stage. An unanticipated outcome of that work was a realization that these same tools could also be used to also trap a wide variety of nomenclatural errors that would otherwise appear in print and in digital form.

Under a Phase II award, we are now extending the scope of data curation and building a framework for extending and distributing N4L information services to users of different classes. The NamesforLife database has undergone further refinement to improve the accuracy of the underlying data. Among the problems that remained at the conclusion of Phase I were a number of long-standing errors that have accumulated in the literature and public databases that were attributable to practices that did not conform with the rules of nomenclature or arose through the continual repropagation of errors in both data sources. Wherever possible, we have reviewed the original published sources to confirm the original observations. Objects in the database now retain the source from which the information was taken from. The NamesforLife data model was also modified to correctly model orthographic corrections, automatically created subspecies names (as per Rule 40d of the International Code of Bacterial Nomenclature) and emendations of taxon descriptions.

Web-based tools have been developed to facilitate data entry and retrieval by NamesforLife curators. In addition to entity editing, there are several ways to query for information

including by name, accession number, and a citation matcher to find references currently in the database. The curatorial tools connect to the database through a set of data validators to help curtail certain data entry errors such as ensuring exemplars are only connected to species/subspecies entities. As of December 2008, the database held records on 11,407 named bacterial and archaeal taxa (8732 species, 491 subspecies, 2184 higher taxa), along with 7747 records identifying verified specimen holdings in biological resource centers and links to 7365 references in which the related taxonomic and nomenclatural acts were effectively and validly published.

The target audience of NamesforLife services is the broad scientific community and others who may need-to-know the precise meaning of biological names or other terms, in correct temporal context as they are encountered in other digital content (scientific or technical literature, regulatory literature, databases, etc). The dynamic, yet asynchronous nature of biological nomenclature and similar terminology poses a significant burden on information providers as they must either invest in constantly maintaining their offerings to keep current or shift that burden to their end-users. If the former, the costs can be significant and, absent a means of synchronizing updates across an entire domain of knowledge, end users are still confronted with apparent discrepancies across data sources and content providers. If the burden is shifted to end-users, they must then locate alternative information sources, typically hosted through a web portal, that must be queried separately. This makes utilization of content cumbersome and can lead to considerable ambiguity.

The NamesforLife approach is to semantically enable content in a manner that is transparent to end-users at two points in the value chain: at the source (the data provider or publisher) and at the client side (the end-user). In either case, the end-user experience is the same. At each occurrence of a validly published bacterial or archaeal name, they can have access to precise authoritative information by simply clicking on the name. Tools to enable publishers content at the prepublishing stage by embedding persistent N4L identifiers into their content ensure that their readers will always have access to the correct meaning of the name (as well as additional information), even if the name has changed since publication. A web-based client has also been developed that supports semantic enablement of other digital content, one-the-fly, providing similar seamless access to NamesforLife content at each point where a name occurs. Collaborative prototyping of these services with early adopters will be discussed.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Phase II STTR Award DE-FG02-07ER86321 A001.

GTL

Standards in Genomic Sciences: an Open-Access, Standards-Supportive Publication that Rapidly Disseminates Concise Genome and Metagenome Reports in Compliance with MIGS/MIMS Standards

Scott H. Harrison[1]* (harris41@msu.edu), Sam V. Angiuoli,[2] Patrick S.G. Chain,[1,3,4,5] Dawn Field,[6] Frank-Oliver Glöckner,[7] Lynette Hirschman,[8] Eugene Kolker,[9,10] Nikos Kyrpides,[4] Susanna-Assunta Sansone,[11] Lynn M. Schriml,[2] Peter Sterk,[6,11] David W. Ussery,[12] Owen White,[2] and **George M. Garrity**[1,5]

[1]Microbiology & Molecular Genetics, Michigan State University, East Lansing, Mich.; [2]Institute for Genome Sciences and Dept. of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, Baltimore, Md.; [3]Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, Calif.; [4]DOE Joint Genome Institute, Walnut Creek, Calif.; [5]Center for Microbial Ecology, Michigan State University, East Lansing, Mich.; [6]NERC Centre for Ecology and Hydrology, Oxford, United Kingdom; [7]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany; [8]Information Technology Center, The MITRE Corporation, Bedford, Mass.; [9]Seattle Children's Hospital Research Institute, Seattle, Wash.; [10]Division of Biomedical and Health Informatics, Dept. of Medical Education and Biomedical Information, University of Washington, Seattle, Wash.; [11]EMBL—European Bioinformatics Institute, Cambridge, United Kingdom; and [12]Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark

Project Goals: The Standards in Genomic Sciences eJournal (SIGS; http://standardsingenomics.org) is a newly established standards-supportive publication being developed to report on the exponentially increasing volume of genomic and metagenomic data. As the first standards-based journal in the life sciences, SIGS will provide immediate open access to its content from the outset, on the principle that making research freely available to the public supports a greater global exchange of knowledge to better accelerate scientific discovery.

The Standards in Genomic Sciences eJournal (SIGS; http://standardsingenomics.org) is a newly established standards-supportive publication being developed to report on the exponentially increasing volume of genomic and metagenomic data[1]. SIGS will also include coverage of detailed standard operating procedures, meeting reports, reviews and commentaries, data policies, white papers and other gray literature that are relevant to genomic sciences, but absent from the scholarly literature. To achieve an objective of standards compliance, SIGS has been designed to use the specification for the minimum information about a genome

sequence (MIGS)[2]. MIGS is developed and maintained by the Genomic Standards Consortium (GSC), a group formed to promote the development of standardized annotations of genomic investigations. As a specification and checklist, MIGS has helped inspire formal models for gathering information routinely included in primary publications such as environmental context, biotic relationship, relationship to oxygen, source material identifiers, nucleic acid sequence and sample metadata as well as the overall sequencing assay and genome annotation protocol. MIGS has been developed to address the genomic information of bacteria, archaea, eukaryotes, plasmids, viruses, and organelles. MIGS has been extended into a specification of minimum information about a metagenome sequence (MIMS). SIGS is the first standards-based journal in the life sciences and will provide immediate open access to its content from the outset, on the principle that making research freely available to the public supports a greater global exchange of knowledge to better accelerate scientific discovery.

SIGS presents new opportunities to tightly integrate biological data and peer-reviewed content. Where the GSC and other standards initiatives have sought to develop interoperable approaches to ensure consistency in semantic and syntactic annotation of genomes and metagenomes[3,4], SIGS will apply and extend these standards to published. While data sets may continue to grow, new goals for research emerge and database structures and applications change, SIGS will provide static, archival snapshots of genomic data and metadata as points of record that are enriched with interpretative commentary and authenticated by peer review and formal validation against specific versions of the defined MIGS/MIMS checklists. The usage of MIGS/MIMS checklists will help guide SIGS as a vehicle for rapidly publishing concise, highly structured short reports of sequenced genomes so that readers can readily make comparisons across taxa, and link out from these comparisons to knowledge existing elsewhere in the literature. Automated scoring will complement the peer review process to address coverage of standards[2] as implemented in minimal and extended forms of XML-based tagging approaches such as can be done for MIGS/MIMS with the Genomic Contextual Data Markup Language (GCDML)[3]. Furthermore, authors may use the opportunity of publishing in SIGS to make further claims concerning their reported data set in terms of how it may be a new finding for a novel clade or niche. As an open access journal, SIGS will be distributed to readers at no cost, with publication costs initially being absorbed through grants from the Michigan State University Foundation and the U.S. Department of Energy (DE-FG02-08ER64707). The estimated throughput for the initial year of publication in 2009 will be 200 articles, with a two to three week peer-review cycle.

Currently, information for over 50 short genome report articles has been provided to SIGS, and archival draft mockups have been marked up in XML based on the with the NLM DTD (http://dtd.nlm.nih.gov) for both short genome reports and short metagenome reports. SIGS has been registered with an ISSN number and has attained CrossRef membership so that digital object identifiers (DOIs) can be assigned to each published manuscript. The editorial workflow and website for SIGS has been deployed with the Open Journal Systems (OJS) software[1], and OJS review forms have been customized to address MIGS/MIMS compliance for some of the organism types that will be reported. Specifically, for the benefit of those contributing content, SIGS has been designed to present authors with example manuscripts based on uploaded annotations of MIGS/MIMS-compliant data during the initial stages of the editorial workflow. This allows for flexibility where authors can begin submission with either weakly or fully populated grids of MIGS/MIMS compliant metadata and update a previously generated document with additional metadata, comments, and insights concerning a reported genome or metagenome. Especially in the case of metagenomics, it is expected that each aspect of an associated biosystem may not fully correspond to the simple content model of the standards specification. Comments by the author can serve to address this for consideration by peer review in the journal. We will develop automated services that will be integrated throughout the editorial workflow to simplify the process of standards compliance and generation of open, searchable, peer-reviewed content.

References

1. Garrity, G. M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S. A., Angiuoli, S., Cole, J. R., Glöckner, F. O., Kolker, E., Kowalchuk, G., Moran, M. A., Ussery, D., and White, O. 2008. Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 12:157-60.
2. Field, D. et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 26:541-7.
3. Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., Glöckner, F. O., and the Genomic Standards Consortium. 2008. A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 12:115-121.
4. Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N., Schriml, L. M., Field, D. and the EnvO Project. 2008. Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata. *OMICS* 12:129-136.

GTL

The Ribosomal Database Project: Tools and Sequences for rRNA Analysis

J.R. Cole* (colej@msu.edu), Q. Wang, B. Chai, J. Fish, E. Cardenas, R.J. Farris, D.M. McGarrell, G.M. Garrity, and **J.M. Tiedje** (tiedjej@msu.edu)

Michigan State University, East Lansing, Mich.

Through its website (http://rdp.cme.msu.edu), the Ribosomal Database Project (RDP) offers aligned and annotated rRNA sequence data and analysis services to the research community (Cole et al., 2008). These services help research-

ers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Updated monthly, the RDP maintained 715,637 aligned and annotated quality-controlled rRNA sequences as of December 2008 (Release 10.6; Fig. 1). The *my*RDP features have grown to support a total of over 2,600 active researchers using their *my*RDP accounts to analyze over 2,000,000 pre-publication sequences in 22,721 sequence groups.

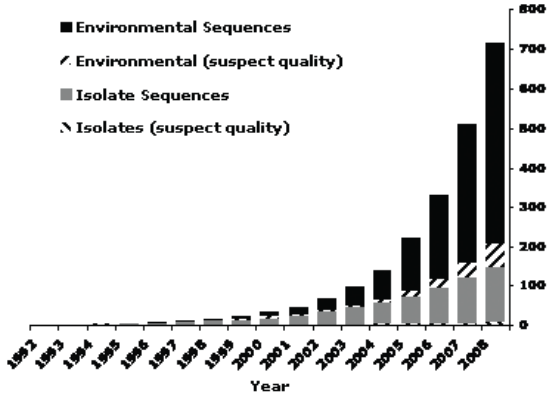


Figure 1. Increase in number of publicly available Bacterial and Archaeal small-subunit rRNA sequences. Suspect quality sequences were flagged as anomalous by Pintail in testing with two or more reference sequences from different publications.

New Bacterial and Archaeal Alignments: A major update to the RDP data sets was released in May 2008. Unlike the RDP 9 series of releases, the RDP 10 series provides an up-to-date aligned and annotated Archaeal data set along with the Bacterial data set. In addition, RDP 10 series alignments are created using the INFERNAL secondary structure based aligner (Nawrocki and Eddy, 2007). Both INFERNAL and RNACAD, our previous aligner, are based on stochastic context-free grammars and provide a high-quality secondary-structure aware alignment. The INFERNAL aligner provides several significant advantages over RNACAD. The INFERNAL aligner is about 25 times faster; it provides a much more intuitive handling of sequencing errors, and solves some known problems with incorrect alignment of short partial sequences. We trained the INFERNAL aligner on a small hand-curated set of high-quality full-length rRNA sequences derived mainly from genome sequencing projects. A relatively small training set of under 1000 representative sequences kept the alignment size small enough for reliable hand-adjustment of homology information in the training set. All *my*RDP users' sequences have been upgraded to the RDP 10 alignments. By adding Archaea, the RDP is responding to the needs of our user community. In addition, this release incorporates recent reevaluations to the *Firmicutes* and *Cyanobacteria* proposed by Bergey's Trust, along with additional published informal taxonomies for the *Acidobacteria*, *Verrucomicrobia*, OP11, and other less-well-studied areas of microbial diversity.

New High-Throughput Pyrosequencing Analysis Pipeline: In May 2008 the RDP released a new pipeline that provides tools to support analysis of next-generation ultra high-throughput rRNA sequencing data. This pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries.

In the initial processing steps, raw sequence reads from multiple samples are sorted using sample-specific tag sequences. The mapping between the tag sequence and sample name is designated in a tag file. Four quality filters can be applied in the initial processing step. For taxonomy-independent alignment, the trimmed reads are aligned using the fast INFERNAL aligner trained on a small, hand-curated set of high-quality full-length rRNA sequences derived mainly from genome sequencing projects. Reads are then clustered into Operational Taxonomic Units at multiple pairwise distances using custom code implementing the complete-linkage clustering algorithm. Specialized tools provide common ecological metrics including: Chao1, Shannon Index and rarefaction. In addition, the processed data can be downloaded in formats suitable for common ecological and statistical packages including SPADE, EstimateS and R. Other options are available to cluster data from multiple samples, to combine alignments, to extract specific sequences from the dataset, to select representative sequences from clustered sequences and to produce comparative metrics among samples.

Several existing RDP tools are used for taxonomy-based analysis. The RDP Classifier provides fast and reliable classification of short sequence reads. The RDP Library Compare program can be used to detect differentially represented taxa between samples, and the RDP Sequence Match tool can be used to find the closest sequences in the RDP database for each sequence in a sample.

New "Taxomatic" Visualization Tool implementing SOSCC: This tool displays a color heat map representation of distances between large sets of sequences. It helps spot errors in the underlying taxonomy. Users can select a set of RDP and/or *my*RDP sequences and display the sequences in taxonomic order in Taxomatic. Thousands of sequences can be visualized at one time, or users can pan and zoom to examine individual sequences. Taxa that are phylogenetically incoherent or are misplaced stand out visually in the representation, as do individual sequences. Any taxonomic group or individual sequence can be highlighted, and tooltips are displayed upon mouseover to display information about individual sequences. The SOSCC alogrithim provides a supervised reclassification of sequences and reordering of the distance matrix. The Taxomatic and SOSCC leverages work originally funded in a separate DOE grant.

New Web Services: The RDP offers new SOAP web services interfaces for the RDP SeqMatch and Classifier tools. These web services have been tested with Java and Ruby. Researchers can incorporate these web services to their own analysis pipelines to make use of these popular RDP tools.

New RDP Class Assignment Generator: This new educational tool for professors or instructors provides a lesson plan

along with individualized material. This generates unique sequences that can be easily distributed to a classroom, providing easy-to-follow instructions for students, and providing an answer key to evaluate the performance of students. The instructor is walked through a simple form that asks about the number of students, the level of difficulty (the number of sequences to be assigned for each student), and other information about the class. The tool then produces a set of sequences for each student modified from existing sequences. The modifications follow evolutionary principles and conserve the rRNA secondary structure. Since each student receives a customized set of sequences, there is less chance of students sharing results. The students are asked to analyze the sequences, and the instructor is provided with a key containing the correct classification for each student's sequences.

RDP User Surveys: Over the last year the RDP implemented a new user survey system to help obtain user input on directions for the RDP. Each survey asks a single question and is displayed for approximately two weeks as an overlay when users access the RDP website. Users may either answer the question or decline to answer before continuing to the RDP website. A browser cookie is used to keep users from seeing the same survey question twice. The results of the RDP surveys are available on the RDP website (http://rdp.cme.msu.edu/misc/surveys.jsp).

References

1. Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje (2008). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* doi:10.1093/nar/gkn879.
2. Nawrocki, E.P. and Eddy, S.R. (2007). Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.,* 3:e56.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-FG02-99ER62848.

Communication

GTL

Genome Management Information System: A Multifaceted Approach to DOE Systems Biology Research Communication and Facilitation

Anne E. Adamson, Shirley H. Andrews, Jennifer L. Bownas, Sharon Burris, Kris Christen, Holly Haun, Sheryl A. Martin, Marissa D. Mills, Kim Nylander, Judy M. Wyrick, Anita J. Alton, and **Betty K. Mansfield*** (mansfieldbk@ornl.gov)

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

Project Goals: Help build the critical multidisciplinary community needed to advance systems biology research for DOE energy and environmental missions and foster industrial biotechnology. The Genome Management Information System (GMIS) contributes to DOE Genomics:GTL program strategies and communicates key GTL scientific and technical concepts to the scientific community and the public. We welcome ideas for extending and improving communications and program integration to represent GTL science more comprehensively.

Concerted communication is key to progress in cutting-edge science and public accountability. With support from the Department of Energy's (DOE) Office of Science, the Genome Management Information System (GMIS) has for 20 years been the main communication resource supporting the Human Genome Project (HGP). However, since 2000 our primary focus has been to help plan, and communicate DOE's Genomics:GTL (GTL, formerly Genomes to Life) program enabled by the HGP. The goal of GTL is to attain a predictive, whole systems–level understanding of microbes and plants to help enable biobased solutions to DOE missions. Our mission is to work with DOE staff and the broad scientific community to communicate biological science challenges and findings to stimulate advances at interdisciplinary interfaces, democratize access to the growing bounty of resources and data, and drive more-informed scientific and societal discourse. Our goals focus on three areas: (1) facilitate GTL planning, research, and communication; (2) respond to communication needs of related projects; and (3) communicate about DOE genomics research and potential applications.

Technical communication integrating all facets of GTL research is critical for spurring innovation at the most rapid pace and at the lowest cost. Such communication is important to achieving DOE missions and, ultimately, fostering U.S. competitiveness through growth in the industrial and environmental sector of the biotechnology industry spurred by DOE genomics research.

Throughout the HGP (1989 to 2003), GMIS strategic networking and communication helped promote collaborations and contributions from numerous fields and reduced duplicative scientific work in the growing genomics community. GMIS staff and the resources we created became the primary "go-to" source for information on all things genomic for much of the scientific world, the media, and the public. A large collection of, informative literature; websites; large-format exhibits; and graphics forms the core of these resources, which are assessed frequently for value, timeliness, and cost-effectiveness. Hundreds of thousands of document copies have been distributed. In addition, GMIS websites annually receive some 20 million page views (224 million hits), many from people who are just learning about genomics and systems biology. Through our resources, networking at various professional scientific and related education meetings, and partnerships, we continue to broaden our reach and focus the attention of those in the national media, government, academia, industry, education, and medicine on DOE genomics and systems biology research.

For the scientific community, communication and research information integration are even more important for GTL than for the HGP, which relied on one dominant technology—DNA sequencing—and produced one major data set—DNA sequence. This new generation of biology is more complex and involves a wider array of technologies, many just emerging, with new types of data sets that must be available to a larger, more diverse research community. Moreover, disparate groups of interdisciplinary scientists must be engaged to achieve the productive dialogue leading to research endpoints that will ensure the success of GTL. The stakes are high: GTL resources and data have the potential to enlarge the research community working on biotechnological approaches to DOE missions, resulting in more rapidly evolving scientific thinking and progress in these and related areas of critical global importance. Communication strategies must be dynamic and evolve along with programmatic needs.

Since 2000, GMIS GTL communication and research integration strategies have included helping facilitate scientific workshops to develop GTL program plans; producing GTL symposia at national scientific meetings; and creating numerous informational resources and tools used by scientists, program administrators, and others. Research plans and reports we have produced with the research community are: *DOE Genomics:GTL Roadmap: Systems Biology for Energy and Environment* (August 2005), *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* (June 2006), and *Carbon Cycling and Biosequestration: Integrating Biology and Climate Through Systems Science* (December 2008). Other work in progress includes reports from two

workshops—on GTL computing (knowledgebase) and on biofuels sustainability as well as this abstracts booklet. We also continuously update and enhance GTL's web presence.

In addition to helping drive communication within the scientific community, GMIS will continue to leverage the high level of public interest in genomic science with our established and future resources to inspire a similar wonder at the challenging new task before us: Learning how genomic "parts" (i.e., genes, regulatory components, and networks) work together to produce the processes of life. GTL pursues this grand scientific challenge via investigations in microbial and plant systems, whose sophisticated biochemical abilities are just now being understood and tapped. We will help communicate the excitement of these investigations and their potential applications within the growing interdisciplinary research community and to broader audiences.

> "Interdisciplinary research…is a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice." [National Academy of Sciences, National Academy of Engineering, and Institute of Medicine of the National Academies, *Facilitating Interdisciplinary Research*, The National Academies Press, Washington, D.C., 2005.]

Ethical, Legal, and Societal Issues

GTL

Intellectual Property and U.S. Public Investments in Research on Biofuel Technologies

Kerri Clark* (kleclark@ucdavis.edu), Rohan Patel, Kyle Jensen, and **Alan Bennett**

Public Intellectual Property Resource for Agriculture, University of California, Davis, Calif.

http://www.pipra.org

Project Goals: We propose to characterize the IP landscape underlying biofuel research and development as it exists globally today. This will entail a systematic analysis and the creation of an IP database in this technology sector which captures both public and private ownership of patents. We anticipate that this "universe" of IP will encompass research in mechanical engineering, chemical engineering, microbiology, fermentation science, biochemistry, genomics and plant sciences. This analysis will provide an assessment of the structure of IP ownership between the private and public sectors, the identification of key private and public sector players in the biofuel sector and the identification of fundamental technologies that may represent barriers for further technological development.

Increased interest in the replacement of fossil fuels with biofuels to combat global warming and increase national security has resulted in a surge in biofuel research whose outcomes are adding to an already complex intellectual property (IP) landscape. An understanding of the biofuel IP landscape can be used to better inform policy makers, sponsors, institutions and researchers to promote and conduct commercially viable research, which will support the maximization of returns on research investments. To increase this understanding we, at the Public Intellectual Property Resource for Agriculture (PIPRA) group, are mapping the IP landscape of biofuel technologies focusing on bioethanol production from cellulosic biomass. This landscape will be used to analyze global patenting activity including identifying the predominant patent applicants, technology advances and geographical patenting trends.

Bioethanol can be produced from lignocellulosic biomass by either enzymatic hyrdolysis (saccharification) of the plants' polysaccharides to sugars and then fermenting the sugars to ethanol; or by converting the lignocellulosic biomass to syn gas and catalytically converting the syn gas to ethanol. We have completed an exhaustive survey of patents on technologies used in lignocellulosic derived bioethanol production via saccharification and fermentation. This survey identified approximately 1400 patents and patent applications (66% saccharification, 26% fermentation and 18% general processes). Of these, approximately 60% are related to enzyme DNA and protein sequences, and enzyme expressing organisms. Overall, the public sector owns approximately 25% of this IP. In every technology category, the private sector is the major IP owner; except in the category "organisms used in fermentation". In this category, over 90% (122) of the patents and applications are owned by the public sector.

The United States is a major source of innovation in the lignocellulosic biomass saccharification sector. The top recipients of patents in the private sector are Genencor, Novozymes and Novo Nordisk; and in the public sector, the Midwest Research Institute, Hebrew University of Jerusalem and the University of Florida. All but two are United States companies and institutions. Genencor owns the largest IP portfolio, however, over 50 assignees are active in this technology arena. Also, approximately 50% of all patents are applied for in the United States reflecting the large numbers of US assignees and the potential return on investment for a successful biofuel due to the massive consumption of transport fuel in the United States. Our preliminary analysis of the IP landscape associated with the saccharification and fermentation of biomass identified a healthy, competitive landscape in which no single company monopolized the IP.

PIPRA is a not-for-profit organization whose objective is to support innovation in public sector agriculture research institutes for commercial and humanitarian uses, by providing a wide range of technical services for improved IP management. These services include the provision of enabling technologies, generation and analysis of IP landscapes, educational services and the facilitation of licensing and material transfer agreements with member institutions. PIPRA comprises 45 institutional members in 14 countries.

GTL

Implications of Alternative Intellectual Property Rights Management Approaches

David J. Bjornstad* (bjornstaddj@ornl.gov) and Amy K. Wolfe (wolfeak@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: This project will study intellectual property rights topics at the DOE's three Bioenergy Centers and will carry out research to explore the general applicability of the research findings for other, related R&D centers. We will produce papers describing intellectual property rights management practices at the Centers including patenting, licensing, and technology transfer activities.

We will also produce papers describing the general applicability of the findings, within the context of the changing patent landscape and efforts by Congress to reform the patent system. Finally, we will generally support the ORNL ELSI SFA on Intellectual Property Rights topics, with initial efforts directed at the ORNL CNMS.

DOE maintains a vigorous technology transfer program as a key element of its IP activities. As part of the management strategy for administering its Bioenergy Centers, DOE has issued a set of Principles to guide IP practices for new findings emanating from the Centers. The Bioenergy Centers, in turn, have responded with Management Plans for implementing the Principles. The Principles and Plans are innovative, far-reaching and represent a departure from past practices. If successful, DOE might consider them for other similar partnered research ventures. For these reasons, and because they are "zero-based," that is, instituted from the initiation of the Bioenergy Centers Program, the Plans offer a unique opportunity to document how they are implemented, how the incentives they embody influence research partners and potential licensees, and the extent to which they could be adjusted for application to other situations.

However, DOE's changing policy for managing intellectual property rights is taking place within an evolving national patenting landscape that may hold special implications for the development of science and technology. Just as advances in gene sequencing technologies led to incentives for genomic researchers to patents human expressed tag sequences a practice that threatened a potential anticommons, wherein large numbers of patents could reduce access to the human genome, trends in patenting may expand the challenges in seeking patent protection and licensing patents that are issued. In general, the breadth of patentable subject matter is increasing and the scope of individual patents is decreasing, both of which lead to larger numbers of patents. Patent licensing strategies are also changing with patent holders and patent seekers each taking into account opportunities to exploit bargaining advantages. Common property rights organizations, such as patent pools or cross-licensing arrangements are being widely explored. This overall phenomenon is sometimes described as a patent thicket

To ensure the broadest possible applicability for our work we are therefore both gathering data and considering concepts to generalize our findings. Over time we intend to explore patenting issues across other DOE R&D institutions in an effort to ensure the best available information is used both in policy making and in technology transfer programs.

The Biofuels Revolution: Understanding the Social, Cultural, and Economic Impacts of Biofuels Development on Rural Communities

Theresa Selfa[1]* (tselfa@ksu.edu), Richard Goe,[1] Laszlo Kulcsar,[1] Gerad Middendorf,[1] and Carmen Bain[2]

[1]Dept. of Sociology, Kansas State University, Manhattan, Kans.; and [2]Dept. of Sociology, Iowa State University, Ames, Iowa

http://www.ksu.edu/sasw/kpc/biofuels/project_doe.htm

Project Goals: The goal of this project is to provide a better understanding of the socio-economic and cultural implications of biofuels development for rural communities, and to contribute to more informed policy development regarding bioenergy.

A new wave of economic growth is currently sweeping across rural communities in the Midwest region of the U.S., fueled by the construction and expansion of ethanol biorefineries and the expansion of biofuel crop production. While the expansion of the biofuels industry promises to bring jobs and economic vitality to rural communities, it is also creating dilemmas for farmers and rural communities in weighing the benefits of income growth and job growth against safety risks, increased pollution, and the potential of overextending water supplies. Presently, there is little empirical knowledge about the social, cultural and economic impacts of biofuels development on rural communities. This research is intended to help fill these lacunae through an in-depth analysis of the social, cultural, and economic impacts of ethanol biorefinery industry on six rural communities in the Midwestern states of Kansas and Iowa. The goal of this project is to provide a better understanding of the socio-economic and cultural implications of biofuels development for rural communities, and to contribute to more informed policy development regarding bioenergy.

Research Questions:

1. To understand how the growth of biofuel production has affected and will affect Midwestern farmers and rural communities in terms of economic, demographic, and socio-cultural impacts.

2. To determine how state agencies, groundwater management districts, local governments and policy makers evaluate or manage bioenergy development in relation to competing demands for economic growth, diminishing water resources, and social considerations.

3. To determine the factors that influence the water management practices of agricultural producers in Kansas and Iowa (e.g. geographic setting, water management institutions, competing water-use demands as well as producers' attitudes, beliefs, and values) and how these

influences relate to bioenergy feedstock production and biofuel processing.

4. To determine the relative importance of social-cultural, environmental and/or economic factors in the promotion of biofuels development and expansion in rural communities.

Research Methodology
We are in the process of analyzing data from the first three case study communities. The comprehensive methodology includes: demographic analysis; in-depth key informant interviews, three focus groups with farmers, ethanol plant workers, and community leaders; a general population opinion survey of community residents; and a content analysis of local newspapers and print media.

Preliminary Findings
Our preliminary findings suggest that communities have mixed sentiments about the biofuels industry. While many believe that ethanol plants have brought jobs and stabilized local populations, they have not led to an increase in population or overall economic growth in their communities. Many expressed anxiety about biofuels competing for water resources with other potential uses, especially in the more arid parts of Kansas. Moreover, they are concerned about the future of the industry and whether their community will eventually be left with a burden if the industry shifts to second generation biofuels production or other alternative energy sources.

For Additional Information:
Project information and research findings will be available at: http://www.ksu.edu/sasw/kpc/biofuels/project_doe.htm

GTL

Analysis of Global Economic and Environmental Impacts of a Substantial Increase in Bioenergy Production

Wallace E. Tyner (wtyner@purdue.edu), Thomas W. Hertel, Farzad Taheripour,* and Dileep K. Birur

Dept. of Agricultural Economics, Purdue University, West Lafayette, Ind.

Project Goals: The goal of this research is to develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. We will build on the unique strengths of GTAP to analyze economic impacts of alternative bioenergy policies at regional and global levels. We will use the TEM model to evaluate the potential for new lands to be brought into production in the wake of biofuel programs, as well as to validate environmental consequences of these policies and check their feasibility from a fundamental bio-geochemical perspective.

Introduction
The global biofuel industry has been experiencing a period of extraordinary growth, fueled (until very recently) by a combination of high oil prices, implementation of ambitious renewable fuel mandates by developed countries (mainly by the US, EU, and Brazil), and government subsidies. This rapid growth has important economic, environmental, and social consequences at a global scale. This research aims to develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. The project is built on the unique strengths of the Global Trade Analysis Project (GTAP) based at Purdue University (Hertel, 1997). The project has begun in 2007 and will be completed over a three-year period from the starting point. To achieve the goals of this project we have introduced production, consumption, and trade of grain based ethanol and biodiesel from oilseeds into the GTAP database. We have successfully extended the GTAP modeling framework to analyze production, consumption and trade of biofuel and their byproducts across the world. Given the importance of land use implications of biofuel production we have also augmented our model with a land use module to accurately depict the global competition for land between food and biofuel. While we continuously work to improve quality and reliability of our database and modeling framework, we have used them to assess the global consequences of the biofuel production from different points of view. This poster reviews major results which we achieved so far.

Research Progress and Primary Results
We have incorporated three explicit biofuels sectors (grain-based ethanol, sugarcane-based ethanol, and vegetable oil based biodiesel, in to the GTAP database (Taheripour et al., 2007). We have extended this database in different directions to properly trace the link among the biofuel, vegetable oil, food, feed, agricultural, and livestock industries. Unlike the earlier version, the new database covers biofuel subsidies and tariffs as well. We plan to introduce producing biofuels from cellulosic materials in the next version of the database.

We extended the GTAP modeling framework to evaluate impact of biofuel production on world agricultural markets (Birur et al., 2007). In this work, we extended the GTAP model to handle production, consumption and trade of biofuels. The paper simulates the biofuel economy during the time period of 2001-2006 and isolates economic impacts of biofuel drivers (such as crude oil price, the US and EU biofuel subsidies, and replacement of MTBE) from other factors at a global scale. In addition, this work calibrates the GTAP parameters for the biofuel economy.

Biofuels from grains and oilseeds are produced in conjunction with other by-products such Dried Distillers Grains with Solubles and oilseed meals. These by-products play an important role in analyzing economic and environmental impacts of biofuel production. We have introduced biofuel byproducts into the GTAP modeling framework and have shown that incorporating biofuel by-products in such analyses considerably alters the results in systematic ways in the face of biofuel policies (Taheripour et al., 2008).

We have examined the impacts of the US and EU biofuel mandates for the world economy and their consequences for the global land use changes (Hertel, Tyner, and Birur, 2008). Unlike earlier papers in this field which have focused on the individual, national impacts of biofuel mandates, we have examined interactions among these policies as well. It shows how the presence of each of these policies and their combination influence global markets and land use around the world. This work evaluates impacts of mandates on production, consumption, exports, and imports of 18 groups of commodities across the world, divided into 18 regions.

In a more recent work, we have examined implications of the biofuels boom for the global livestock industry (Taheripour, Hertel and Tyner, 2008). We show that the US and EU biofuel mandates will encourage crop production in both biofuel and non biofuel producing regions, while reducing livestock production in most regions of the world. This work indicates that, the non-ruminant industry curtails its production more than other livestock industries. We also show that the biofuel mandates reduce food production in most regions while they increase crude vegetable oils in almost all regions. We finally conclude that, while biofuel mandates have important consequences for the livestock industry, they do not harshly curtail these industries.

This poster highlights major findings of these research activities. It also presents our next steps towards the projects goals. In the next steps we will extend our database and model to incorporate cellulosic ethanol into our framework, evaluate impacts of biofuel policies on global poverty, and finally combine results from GTAP and other branch of the project which aims to extend the Terrestrial Ecosystem Model (TEM) to evaluate environmental consequences (such as greenhouse gas emissions and water use) of alternative bioenergy policies.

References

1. Birur, D., T. Hertel and W. Tyner. 2007. "Impact of Biofuel Production on World Agricultural Markets: A Computable General Equilibrium Analysis." GTAP Working Paper No 53, Center for Global Trade Analysis, Purdue University, West Lafayette, IN.
2. Hertel, T. W. 1997. "Global Trade Analysis, Modeling and Applications," Cambridge University Press, Cambridge.
3. Hertel, T., W. Tyner, and D. Birur. 2008. "Biofuels for all? Understanding the Global Impacts of Multinational Mandates," GTAP Working Paper No. 51, Center for Global Trade Analysis, Purdue University, West Lafayette, IN, USA.
4. Taheripour, F., D.K. Birur, T.W. Hertel, and W.E. Tyner. 2007. "Introducing Liquid Biofuel into the GTAP Database." GTAP Research Memorandum No. 11. Center for Global Trade Analysis, Purdue University, West Lafayette, IN, USA.
5. Taheripour, F., T.W. Hertel, and W.E. Tyner. 2008. "Biofuels and their By-Products: Global Economic and Environmental Implications." Department of Agricultural Economics, Purdue University, West Lafayette, IN, USA.
6. Taheripour, F., D.K. Birur, T.W. Hertel, and W.E. Tyner (2007) 'Introducing Liquid Biofuels into the GTAP Database' GTAP Research Memorandum. Center for Global Trade Analysis, Purdue University, West Lafayette, IN47907, USA.
7. Taheripour, F., T. Hertel, and W Tyner. 2008. "Implications of the Biofuels Boom for the Global Livestock Industry: A Computable General Equilibrium Analysis," Background paper for the 2009 State of Food and Agriculture (SOFA)From the Food and Agriculture Organization of the UN (FAO), Center for Global Trade Analysis, Purdue University, West Lafayette, IN, USA.

GTL

Issues Associated with the Transition from Science to Application in a Fundamental Nanoscience User Center

Amy K. Wolfe* (wolfeak@ornl.gov) and David J. Bjornstad (bjornstaddj@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tenn.

Project Goals: The overarching research question for this task is how best to integrate ELSI scholarship into the conduct of fundamental research and development at the CNMS, thereby enhancing the organization's knowledge of and ability to address ELSI issues before they erupt into potentially polarizing controversies. More specific research- and process-oriented questions this task seeks to answer are the following: 1) Which key stages in nanophase materials R&D have specific linkages with downstream ELSI issues?; 2) Which ELSI-related issues should scientists and science managers be prepared to address, when interacting with different audiences?; 3) What kinds of institutional issues/barriers can be anticipated as the CNMS evolves?; 4) How can ELSI become part of the CNMS "culture" rather than an administrative requirement?; and 5) How can ELSI research and results be made available to members of society as CNMS research and public awareness, knowledge, and concerns evolve over time?

The overarching goal of this project is to integrate Ethical, Legal, and Societal Implications, or "ELSI," into the Center for Nanophase Materials Sciences (CNMS) at Oak Ridge National Laboratory and, by extension, the larger nanoscience and nanotechnology R&D community. ELSI encompasses the wide array of studies and activities that investigate the "other," non-technological side of the emerging science and technology (S&T). While S&T is oriented toward accomplishing challenging, grand, and sometimes poorly understood outcomes ultimately aimed at enhancing societal well-being, ELSI research can be thought of as studying the interactions between emerging S&T and society as they co-evolve, and predicting the implications and consequences of those interactions on science and society.

ELSI research identifies a suite of concerns about the *potential* positive and negative impacts of emerging S&T on soci-

ety. It also isolates the choices that influence the (positive or negative) impacts *actually* experienced by individuals, subsets of society, or societal institutions. ELSI activities may seek to anticipate societal responses that will follow from earlier choices so as to help sensitize early decision makers about the potential implications of their decision-making processes and resulting decisions.

A major challenge for this project is to conduct ELSI research and related activities in a manner that exploits opportunities presented by the project's close proximity to ongoing CNMS R&D. Our goals are to: (a) sharpen the relevance of ELSI inquiry for the R&D community; and (b) enhance the ability of R&D community to address ELSI issues before they escalate into potentially polarizing controversies. This poster summarizes the implications of one key transition in the nanoscience and nanotechnology life cycle—that of the transfer of fundamental science to explicit downstream applications—and the feedback to S&T resulting from this transfer. The poster describes interim results of interviews with CNMS researchers, highlighting their perspectives on the linkage between fundamental R&D and its downstream uses (demonstration, deployment, decommissioning, and disposal) and their insights into societal issues that should be raised in studying this transition.

Acknowledgement:
This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

GTL

Global Net Primary Production and Bioenergy Potentials and the Environmental Consequences: An Analysis with a Process-Based Terrestrial Ecosystem Model

Qianlai Zhuang (qzhuang@purdue.edu), Kai Xu, and Yanyu Lu*

Dept. of Earth & Atmospheric Sciences and Dept. of Agronomy, Purdue University, West Lafayette, Ind.

http://www.purdue.edu/eas/ebdl/

Project Goals: To develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. The goal of this research is to develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. We will build on the unique strengths of GTAP to analyze economic impacts of alternative bioenergy policies at regional and global levels. We will use the TEM model to help develop the land supply curves and to validate environmental consequences of these policies and check their feasibility from the environmental and land use perspectives.

Our project, entitled "Analysis of Global Economic and Environmental Impacts of a Substantial Increase in Bioenergy Production", is designed to develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. The goal of this research is to develop realistic assessments of the economic and environmental impacts of regional and global policies designed to stimulate bioenergy production and use. We will build on the unique strengths of GTAP to analyze economic impacts of alternative bioenergy policies at regional and global levels. We will use the TEM model to help develop the land supply curves and to validate environmental consequences of these policies and check their feasibility from the environmental and land use perspectives.

To date, using a process-based biogeochemistry model, the Terrestrial Ecosystem Model (TEM), we have evaluated the global net primary production (NPP) in natural ecosystems as a way to assess global biofuel potentials. We also evaluated the global agricultural NPP with the ecosystem model. NPP is the net amount of carbon captured by land plants through photosynthesis each year considering the effects of atmospheric climate and chemistry changes. We further evaluated the global net ecosystem production (NEP), which is the net carbon exchange between the terrestrial ecosystems and atmosphere. The changes of global NPP and water used during photosynthesis were also evaluated to assess the importance of freshwater in support of NPP production. In addition, we evaluated global emissions of another potent greenhouse gas, nitrous oxide from both natural and agricultural ecosystems, which is accompanied with the production of NPP. As a next step, we will use this ecosystem and biogeochemistry modeling framework to evaluate the alternative bioenergy policies on environmental consequences such as greenhouse gas emissions and water use.

Appendix 2: Websites

Program Websites

- Genomics:GTL: http://genomicsgtl.energy.gov
- This book: http://genomicsgtl.energy.gov/pubs/2009abstracts/
- *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* http://genomicsgtl.energy.gov/biofuels/b2bworkshop.shtml
- DOE Microbial Genome Program: http://microbialgenomics.energy.gov
- DOE and USDA Biomass Genomics Research http://genomicsgtl.energy.gov/research/DOEUSDA/

Bioenergy Research Center Websites

- Overview http://genomicsgtl.energy.gov/centers/
- BioEnergy Science Center: http://www.bioenergycenter.org
- Great Lakes Bioenergy Research Center: http://www.greatlakesbioenergy.org
- Joint BioEnergy Institute: http://www.jbei.org

Project and Related Websites

- DOE Joint Genome Institute: http://jgi.doe.gov
- MAGGIE: http://masspec.scripps.edu/maggie/
- MicrobesOnline: http://microbesonline.org
- Microbial Protein-Protein Interaction Database: http://MiPPI.ornl.gov
- Public Intellectual Property Resource for Agriculture http://www.pipra.org/
- Virtual Institute for Microbial Stress and Survival: http://vimss.lbl.gov
- Wheat Genetic and Genomic Resources Center http://www.ksu.edu/wgrc

Author Index

Indexed by Page Number

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Institutional Index

Indexed by Page Number