Genomics of Energy & Environment

Fourth Annual
DOE Joint Genome Institute
User Meeting

Sponsored By

U.S. Department of Energy
Office of Science

March 25–27, 2009

Walnut Creek Marriott

Walnut Creek, California

Contents

DOE Joint Genome Institute

2009 JGI User Meeting

March 25-27, 2009

Walnut Creek Marriott, Walnut Creek, CA

All functions will be held at the Walnut Creek Marriott unless otherwise noted.

AGENDA

WEDNESDAY, March 25

Start Time	*End Time*	*Subject*	*Session Chair/Speaker*
9:00 AM	**12:00 PM**	**Workshops**	
		IMG Tutorial	Nikos Kyrpides
		Eukaryotic Annotation Tools	Igor Grigoriev
		JGI 101 - An Introduction to JGI	Jim Bristow
12:00 PM	**1:00 PM**	***Lunch provided for WORKSHOP PARTICIPANTS ONLY on this day***	
1:00 PM	**4:30 PM**	**Plants, Biomass Degradation and Biofuels**	Jim Bristow, Chair
1:00 PM	1:15 PM	Introduction	Eddy Rubin, JGI
1:15 PM	1:45 PM	Pathogen virulence and the plant immune system	Jeff Dangl, U. North Carolina
1:45 PM	2:15 PM	Reverse-engineering cellulosomes in Clostridia	Jamie Cate, EBI
2:15 PM	2:45 PM	Algal engineering for the production of an energy crop	Mike Mendez, Sapphire Energy
2:45 PM	3:15 PM	***Break***	

3:15 PM	3:45 PM	Evolution of Ancient Agriculture in Ants	Cameron Currie, U. Wisconsin
3:45 PM	4:15 PM	*Arabidopsis* Epigenetics	Joe Ecker, Salk Institute
4:15 PM	4:45 PM	New Developments in High Throughput Proteomics Measurements: Augmenting Genomics	Dick Smith, PNNL
5:30 PM	**6:15PM**	**Keynote Presentation**	**Chris Somerville, EBI**
6:15 PM	**10:00 PM**	**Opening Reception and Poster Session**	

THURSDAY, March 26

Start Time	*End Time*	*Subject*	*Session Chair/Speaker*
9:00 AM	**12:00 AM**	**Genome Evolution and Adaptation – Part 1**	Dan Rokhsar, Chair
9:00AM	9:30 AM	Whole Genome Comparisons of *Bacillus subtilis*	Ashlee Earl, Harvard University
9:30 AM	10:00 AM	Genomic Encyclopedia of Bacteria and Archaea	Jonathan Eisen, U. California, Davis
10:00 AM	10:30 AM	***Break***	

10:30 AM	11:00 AM	Fungal Genomics	James Galagan, Broad Institute
11:00 AM	11:30 AM	Genomic Analysis of Adaptation and Speciation in *Mimulus guttatus*	John Willis, Duke University
11:30 AM	12:00 AM	Diatom Comparative Genomics	Ginger Armbrust, U. Washington
11:30 AM	1:00 PM	***Lunch***	
12:00 PM	**1:30 PM**	**Business Meeting-** ***Open Forum with JGI Management***	
1:45 PM	**5:00 PM**	**New Technologies and Functional Genomics**	Len Pennacchio, Chair
1:45 PM	**2:30 PM**	**Keynote Presentation**	**George Church, Harvard University**
2:30 PM	3:00 PM	Pacific Biosciences Sequencing	Steve Turner, Pacific Biosciences
3:00 PM	3:30 PM	***Break***	
3:30 PM	4:00 PM	Short Read Assembly	Pavel Pevzner, UCSD
4:00 PM	4:30 PM	Phylochip Analysis of Microbial Diversity	Gary Andersen, LBNL
4:30 PM	5:00 PM	*Chlamydomonas* Transcriptomics	Matteo Pellegrini, UCLA
5:00 PM	6:00 PM	***Travel to JGI – Bus Service from Marriott***	
6:00 PM	**9:00 PM**	**Reception, Poster Session and Tours at JGI**	
9:00 PM	10:00 PM	***Travel from JGI – Bus Service to Marriott***	

FRIDAY, March 27

Start Time	*End Time*	*Subject*	*Session Chair/Speaker*
8:30 AM	**12:00 PM**	**Genome evolution and adaptation – Part 2**	Eddy Rubin, Chair
9:00 AM	9:30 AM	Population Genomics and the Bacterial Species Concept	Peg Riley, U. Massachusetts
9:30 AM	10:00 AM	Systems Biology in *Caulobacter*	Lucy Shapiro, Stanford Univ.
10:00 AM	10:30 AM	***Break***	
10:30 AM	11:00 AM	Diversity Generating Retroelements	Jeff F. Miller, UCLA
11:00 AM	11:30 AM	To Refseq or Reseq Oryza - That is the Question	Rod Wing, U. Arizona
11:30 PM	**12:15 PM**	**Keynote Presentation**	**Craig Venter, JCVI**
	12:15 PM	***End of User Meeting***	

Speaker Presentations

Abstracts alphabetical by speaker

Phylochip Analysis of Microbial Diversity

Gary Andersen

Lawrence Berkeley National Laboratory, Berkeley, California

Diatom Comparative Genomics

Ginger Armbrust

University of Washington, Seattle

Cellulosome Engineering

Jamie Cate

Joint BioEnergy Institute (JBEI)

Evolution of Ancient Agriculture in Ants

Cameron R. Currie (currie@bact.wisc.edu)

Great Lakes Bioenergy Research Center and Department of Bacteriology, University of Wisconsin, Madison

Host-microbe symbioses have played a critical role in the evolution of biological diversity and complexity. A paradigmatic example is the fungus-growing ant-microbe symbiosis, where coevolution between ants and their microbial symbionts has culminated in one of the dominate herbivores of the Neotropics; the leaf-cutter ants. The ants carefully tend their fungal mutualist, providing it with optimal conditions for growth, and in exchange, the fungus serves as the main food source for the colony. The origin of this mutualism was more than 45 million years ago, and the subsequent shared evolutionary history has generated significant diversity in both the ants and fungi. Recent work has shown that the gardens of fungus-growing ants are host to a specialized, virulent, and coevolved fungal pathogens in the genus *Escovopsis*. To help deal with the garden pathogen, the ants have a mutualistic association with actinobacteria, which produce antibiotics that suppress the growth of *Escovopsis*. In the leaf-cutters, the most derived members of fungus-growing ants, workers use fresh leaf material as substrate for growing their fungal crop. A mature leaf-cutting ant nest can contain >5 million workers, which can collect nearly 400 kg (dry weight) of plant material per year. Interestingly, the fungal mutualist appears to lack the metabolic capacity to degrade most of the recalcitrant plant polymers. Thus, our understanding of the fundamental component of the leaf-cutting ant system, plant breakdown into energy for the ants, is largely unknown. In collaboration with the JGI, we are characterizing the microbial community and its metabolic potential within the fungus

garden of leaf-cutting ants to determine the role bacteria play in contributing to the breakdown of cellulose within this system. An understanding of how recalcitrant plant biomass is converted into energy in this natural and highly-evolved lignincellulose-rich system has the potential to lead to the discovery of novel enzymes and/or microbes relevant to cellulosic ethanol production.

Plant Genomics Roadmap

Jeff Dangl

University of North Carolina, Chapel Hill

Whole Genome Comparisons of *Bacillus subtilis*

Ashlee M. Earl[1] (ashlee_earl@hms.harvard.edu), Florian W. Fricke,[2] Richard Losick,[3] Roberto Kolter,[1] and Jacques Ravel[2]

[1]Harvard Medical School, Boston, Massachusetts; [2]University of Maryland, Baltimore; and [3]Harvard University, Cambridge, Massachusetts

For decades, the bacterial species *Bacillus subtilis* has been a workhorse for the study of molecular genetics and bacterial development. As with most model organisms, there has been a trade-off, sacrificing breadth for depth, as work with the species has focused almost exclusively on one strain, *B. subtilis* 168. Although an unparalleled understanding of the inner workings of *B. subtilis* 168 has been gained, we now recognize that the genotypic and phenotypic potential of the species cannot be accounted for by the characterization of a single strain. We have recently sequenced and compared the genomes of two additional strains of *B. subtilis* to that of *B. subtilis* 168. The comparisons have revealed that the genomes of this species are highly mosaic. Each strain harbors a large number of unique regions that are interspersed throughout a highly conserved and syntenic genomic backbone. The strain-specific regions vary enormously in size (<1 to >100 kilobases) and collectively contain hundreds of genes, a subset of which likely confer different ecological adaptation. These unique regions also contain genes that function in well-characterized developmental pathways (e.g., sporulation and competence) suggesting that these pathways are plastic and, perhaps, subject to environmental selection. While the mechanisms responsible for the appearance of these strain-specific regions are not entirely known, phages undoubtedly play a role.

Arabadopsis Epigenetics

Joe Ecker

Salk Institute, La Jolla, California

Genomic Encyclopedia of Bacteria and Archea

Jonathan Eisen

University of California, Davis

Fungal Genomics

James Galagan

Broad Institute, Cambridge, Massachusetts

Algal Engineering

Mike Mendez

Sapphire Energy

Diversity Generating Retroelements

Jeffrey F. Miller

University of California, Los Angeles

Chlamydomonas Transcriptomics

Matteo Pellegrini

University of California, Los Angeles

Short Read Assembly

Pavel Pevzner

University of California, San Diego

Population Genomics and the Bacterial Species Concept

Margaret Riley (riley@bio.umass.edu)

University of Massachusetts, Amherst

The importance of horizontal gene transfer (HGT) in bacterial evolution has been elevated to such a degree that many bacteriologists now question the very existence of bacterial species. If gene transfer is as rampant as comparative genomic studies have suggested, how could bacterial species survive such genomic fluidity? And yet, most bacteriologists recognize and name as species, clusters of bacterial isolates that share complex phenotypic properties. The Core Genome Hypothesis (CGH) has been proposed to explain this apparent paradox of fluid bacterial genomes associated with stable phenotypic clusters. It posits that there is a core of genes responsible for maintaining the species-specific phenotypic clusters observed throughout bacterial diversity and argues that, even in the face of substantial genomic fluidity, bacterial species can be rationally identified and named. As multiple whole genome sequences become available for more putative bacterial species, the CGH can finally be put to the test. Will it survive?

Systems Biology in *Caulobacter*

Lucy Shapiro (Shapiro@stanford.edu)

Department of Developmental Biology, Stanford University School of Medicine, Stanford, California

We use a systems biology approach to define the complete genetic circuitry that coordinates cell differentiation as a function of a bacterial cell cycle. Microarray analysis of the 3767 genes in the *Caulobacter* genome revealed that the transcription of approximately 15% of the genome is cell cycle controlled. Genes involved in a given cell function are activated at the time of execution of that function. A single regulatory factor, the CtrA response regulator, directly regulates 95 cell cycle genes. CtrA and a second global regulator, GcrA (regulating about 100 genes), oscillate during the cell cycle, controlling the temporal expression of functional modules during cell cycle progression. These events are controlled and coordinated by differential DNA methylation, regulated proteolysis and phosphorylation signaling cascades. Both chromosomal regions and regulatory protein complexes exhibit actin-dependent dynamic localization that is important for cell cycle control. Thus, deciphering the entire regulatory network requires the coordination of multiple levels of control, including the integration of three-dimensional information.

New Developments in High-Throughput Proteomics Measurements: Augmenting Genomics

Richard D. Smith (rds@pnl.gov)

Pacific Northwest National Laboratory, Richland, Washington

Capabilities for broad, comprehensive, and quantitative proteomics measurements are advancing rapidly, and can now be applied to essentially any biological system. To this point the wide-spread application has been limited by proteomics measurement throughput, as well as compromises with the sensitivity, coverage, and quantitative aspects of these measurements. This presentation will outline the insights obtainable from proteomics measurements and how they augment genomics, e.g. by improving gene definition and annotation, and understanding function using information from protein localization, modification states, and activity-based proteomics approaches. Additionally, a new platform will be described that is based upon much faster separations combined with mass spectrometry, that broadly addresses these issues, and provides the basis for the much broader use of proteomics.

Pacific Biosystems Sequencing

Steve Turner

Pacific Biosystems, Menlo Park, California

Genomic Analysis of Adaptation and Speciation in *Mimulus guttatus*

John H. Willis (jwillis@duke.edu)

Duke University, Durham, North Carolina

Combining ecological and genomic experimental approaches allow us to address long-standing questions about the origins of biological diversity. How do populations adapt to complex and often unpredictable environments? How does adaptation lead to the formation of new species? What is the molecular genetic basis of adaptation and speciation at the whole-organism level? The wildflower *Mimulus guttatus* and its closely related species display tremendous genetic variation in traits related to adaptation to diverse and extreme environments. Because of their remarkable ecological diversity, and the ease with which they can be crossed and studied in nature and the lab, ecologists and evolutionary geneticists have intensively studied these plants for over 60 years. Now scientists at JGI, in collaboration with the Mimulus research community, have completed a draft genome sequence of *Mimulus guttatus*, the first asterid (one of the two main groups of eudicots) to be sequenced. This reference genome, in conjunction with additional genomic tools developed with support from NSF, is already enabling the discovery of genes involved in plant environmental adaptation and speciation. In this talk I will focus on our progress towards identifying genes underlying key developmental, physiological, and reproductive traits related to adaptation to seasonal water availability, salt stress, and the toxic soils of mine tailings.

To Refseq or Reseq Oryza—That is the Question

Rod A. Wing (rwing@ag.arizona.edu)

Arizona Genomics Institute, University of Arizona, Tucson

Genome sequencing technology is advancing at a rapid pace with promises of *de novo* whole genome sequences for the majority of crop and model plants within the next few years. How these genomes will be generated depends on the biological questions that they will be used to address.

Our consortium developed a set of genus-level genomics resources to address questions in plant evolution, under the rubric—the *Oryza* Map Alignment Project (OMAP). *Oryza* is composed of 2 domesticated rice species and 22 wild relatives with wide geographical distribution and habitats, 10 genome types (6 diploid, 4 polyploid), and a 3.6 fold genome size variation. The OMAP resource is comprised of a set of BAC-based physical maps from 13 *Oryza* species aligned to the IRGSP rice RefSeq, and a set of five Chr3 short arm RefSeqs.

Using these resources we made the following key discoveries: 1) Analysis of structural variation between the rice RefSeq and its AA and BB genome relatives revealed that despite extensive observed colinearity, each genome is in "flux" and can differ by at least 40-50Mb (>10% of total genome size); 2) Large-scale sequence analyses of four biologically important regions (*Adh1*, *MOC-1*, *Hd-1*, and a recombinational hotspot on Chr3) revealed dynamic evolution of the *Oryza* genomes within the last 15 MY fashioned by a variety of lineage-specific structural rearrangements. More surprisingly, we uncovered a striking flux in gene content even among the most closely related species (AA and BB genomes). 3) Analysis of the repetitive fraction of the *Oryza* enabled the discovery

of an ancient retrotransposon family 'RWG' which was shown to be responsible for the greater than two-fold genome size variation found in the diploid species *O. granulata* [GG] and *O. australiensis* [EE].

Given that even the most closely related *Oryza* species contain significant differences in genomic content and organization, we developed a cost-effective method to generate high-quality *de novo* reference sequences of large plant genomes by combining "old school" physical maps with next generation sequencing technology. As a pilot, we sequenced the 18Mb Chr3S from *O. barthii* using 6 pools of 28 BACs (each pool ~3Mb) selected from the physical map. Each pool was sequenced with 454 Titanium and GS FLX paired end chemistries. The result was a Chr3 arm sequence with a contigN50 of 14.3kb, a scaffoldN50 of 3.16Mb, and which had 90% of its length present in just 6 of 44 scaffolds—the largest of those being over 6Mb in length. Nucleotide accuracy was assessed by comparing overlap regions between neighboring pools and resulted in an error rate of only 2.2bp per 10kb. To assess the utility of the sequence for gene identification, we mapped known *O. sativa* genes onto the *O. barthii* Chr3S assembly. Of the 3,127 genes identified in the sequence, 2,333 (75%) were completely covered, and 92% had greater than 90% coverage along their length.

Based on these key biological findings and technological advancements, we argue that it is critical to first generate high-quality RefSeqs for all 8 AA, and 1 representative of the 9 other *Oryza* genome types, using a NextGen/BAC Pool approach (or another TBD) and then utilize re-sequencing methods (i.e. Illumina or SOLID) to capture species-specific allelic variation.

Poster Presentations

Posters alphabetical by first author. *Presenting author

Development of a Method of Measuring the Activity Laccase Segregated by Monokarions of *Pleurotus ostreatus* During the Process of Biological Pretreatment of Lignocellulose for the Production of Bioethanol

Manuel Alfaro* (manuel.alfaro@unavarra.es), Lucía Ramírez, and Antonio G. Pisabarro

Universidad Pública de Navarra, Pamplona, Spain

The utilization of low-value substrates such as lignocellulosic wastes offers a great potential for reducing the production costs of bioethanol. The biological process of bioethanol production using lignocellulose as feedstock requires its delignification to liberate the cellulose and hemicellulose from their complex with lignin. *P. ostreatus* produces various lignolytic activities that act in this process. One of them is the laccase. In this work, we have set a protocol for measuring the laccase activity secreted in solid cultures performed on poplar (*Populus alba*) sawdust, we have followed its evolution over the culture time and we have studied its variation in different strains.

We show here (1) that the extractable laccase activity was higher when the cultures were inoculated with mycelium resuspended in water than when the mycelium was resuspended in complex media, (2) that different monokaryotic strains displayed different profiles of laccase activity over the culture time, and (3) that there is evidence suggesting the presence of an inducer of the secretion laccase activity in the sawdust that can be extracted from the wood. Moreover, monokaryons mk009 and mk116 displayed the highest secreted laccase activity and they are, therefore, the best suited ones for the development of a biological pretreatment of lignocellulosic substrates for the industrial production of bioethanol. These two monokaryons showed different patterns of temporary expression of laccase activity. Finally, monokaryon mk009 produces laccase activity earlier in the culture and maintains this activity over all the culture time analyzed.

Metagenomic Characterization of Compost and Rain Forest Soil Microbial Communities

Martin Allgaier[1,3]* (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] Blake Simmons,[1,6] Terry C. Hazen,[1,5] and Phil Hugenholtz[1,3]

[1]Joint BioEnergy Institute, Emeryville, California; [2]University of California, Davis; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Lawrence Livermore National Laboratory, Livermore, California; [5]Lawrence Berkeley National Laboratory, Berkeley, California; and [6]Sandia National Laboratory, Livermore, California

Microorganisms are a promising source of novel carbohydrate-active enzymes (cellulases, hemicellulases, and ligninases) 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 tool MG-RAST. 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.

Genome Sequencing of a Multidrug-Resistant *Salmonella* Typhimurium Isolate in Hong Kong

Chun Hang Au[1]* (tommyau@cuhk.edu.hk), Winnie Wing Yan Chum,[1] Patrick Tik Wan Law,[1] Ka Hing Wong,[1] Julia Mei-Lun Ling,[1] Kai Man Kam,[2] Yin-Wan Wendy Fung,[1] and Hoi Shan Kwan[1] (hskwan@eservices.cuhk.edu.hk)

[1]The Chinese University of Hong Kong, Hong Kong SAR, PR China; and [2]Department of Health, Hong Kong SAR, PR China

Food borne diseases such as salmonellosis are common public health issues around the world. There are an estimated 1.4 million non-typhoidal salmonellosis and 15,000 hospitalizations annually in the United States alone. *Salmonella* Typhimurium is one of the most important serotypes associated with the infections. The emergence of multidrug-resistant (MDR) *S.* Typhimurium complicates the treatment and poses a serious threat to the public. Revealing the repertoire of drug resistance determinants is the key to accurately identify and track the MDR strains. We are sequencing the genome of a Hong Kong clinical isolate, which exhibits a wide-spectrum resistance to a panel of antibiotics. Almost three-fold genome coverage of sequencing reads are mapped to 3 sequenced strains (LT2, DT104, and SL1344) and revealed 218-655 possible polymorphisms between the MDR isolate and the 3 strains. Unmapped reads (1.5 Mb) suggest the presence of laterally transferred elements acquired by the MDR isolate. Further sequence characterization will also be presented.

A Genomic Analysis of *Thermus aquaticus* and *Thermus brokianus*

Frank O. Aylward,[1] **Garret Suen**[1,2]* (gsuen@wisc.edu), Cameron R. Currie,[1,2] Susana F. Delano,[2] Jean F. Challacombe,[3] Thomas Schoenfeld,[4] David Mead,[4] and Phil Brumm[5]

[1]Department of Bacteriology, University of Wisconsin, Madison; [2]Great Lakes Bioenergy Research Center, Madison, Wisconsin; [3]DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico; [4]Lucigen Corporation, Middleton, Wisconsin; and [5]C56 Technologies, Middleton, Wisconsin

Superheated thermal pools are among the most hostile environments known. Extremes of temperature and pH, low nutrient availability, and toxic heavy metals and gases select for organisms that have adapted novel lifestyles and unique biochemistries to survive in these harsh conditions. We have been sampling thermal pools from Yellowstone National Park in search of organisms and enzymes that can be used to remove bottlenecks in the breakdown of recalcitrant biomass and conversion into biofuels. Here we present a biological characterization of two bacteria in the genus *Thermus*: *Thermus aquaticus* and *Thermus brockianus*. Fluorescent microscopy of these organisms shows the appearance of unusual spherical bodies when grown under micro-aerobic conditions. These morphologies may have implications for novel pathways that can be developed into tools for biofuels based feedstocks. To further understand the biology of these two bacteria, we performed whole-genome sequencing with the JGI, and present a preliminary analysis of these draft genomes including a COG category analysis, a taxonomic distribution analysis of their proteins, and a comparative metabolic reconstruction analysis. Taken together, these data indicate that these organisms are highly specialized for their particular niche and likely contain proteins and enzymes that are of valuable for bioenergy.

Targeting Bioenergy, Environment and New Protein Families: Community Nominated JGI/MCSG Structural Genomics Pilot Project

G. Babnigg[1]* (gbabnigg@anl.gov), C.A. Kerfeld,[2] and A. Joachimiak[1]

[1]Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, Illinois; and [2]DOE Joint Genome Institute, Walnut Creek, California

The Joint Genome Institute generates sequence information at a constantly accelerating rate for a list of species relevant to DOE missions, namely, species affecting global carbon cycling, microbial communities or single species that play a role in the degradation of lignocellulosic material, species with rich metabolic potential, as well as highly diverse microbial species sequenced as a part of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. The functional characterization of some of the newly uncovered gene families, especially those without significant sequence homology to existing genes, is one of the roadblocks in this modern high-throughput science. A 3D structure does not require any a priori knowledge about the protein, contributes new data and can jumpstart functional assignment and assay development by providing an atomic level 3D description of a given protein along with the different gene expression clones and purified proteins.

The Midwest Center for Structural Genomics (MCSG) as part of the Protein Structure Initiative (PSI) provides structural coverage of major protein superfamilies with granularity allowing 3D homology modeling of a large number of proteins using only computational methods. The ultimate goal of PSI is to build a foundation for 21st century

structural biology where the structures of virtually all proteins will be found in the Protein Data Bank (PDB) or derived by computational methods. The structural genomics high-throughput pipeline developed at MCSG comprises: (1) classifying all available genomic sequences to establish a prioritized target set of proteins, (2) cloning and expressing proteins of microbial and eukaryotic origin, (3) purifying and crystallizing native and derivatized proteins for X-ray crystallography, (4) collecting data and determining structures using synchrotron sources, (5) analyzing structures for fold and function assignment, and homology modeling of related proteins. The structural genomics pipeline takes advantage of significant advances in molecular and structural biology including synchrotron facilities, dedicated PX beamlines, advanced software and computing resources. The structural genomics technologies can be applied to a wide range of protein targets and are very well suited for proteins originating from microbial communities. The MCSG targets are selected from large protein families with no structural representative, biomedically important pathogens and higher eukaryotes, metagenomics projects and also include the scientific community nominated targets.

Recently the Joint Genome Institute (JGI) User Community and JGI scientists nominated a large set of targets from more than 50 microbial species. Nearly 70 scientists proposed more than 700 targets for structure determination. This poster will summarize the results from the first batch of targets processed to date. Some of these proteins are targeted to bacterial micro-compartments and the structural biology provides an in-depth insight into the function and potential mechanism of action of these proteins.

This work was supported by NIH Grant GM074942 and by the U.S. DOE, OBER contract DE-AC02-06CH11357.

Mining Phytopathogen Genomes for Enzymes and Secretion Systems

Venkatesh Balakrishan,* Jeremy Glasner (jglasner@wisc.edu), and Nicole Perna

University of Wisconsin, Madison

Effective fermentation of lignocellulosic biomass requires improved methods for hydrolysis of cellulose and hemicellulose. Many phytopathogenic organisms are capable of degrading and utilizing these compounds and may be a source of novel enzymes and pathways for lignocellulosic ethanol production. The bacterial family Enterobacteriaceae, that includes the well-known *E. coli* species, also includes 5 genera of plant pathogens (*Erwinia, Pectobacterium, Dickeya, Pantoea* and *Brenneria*). We have sequenced genomes from representatives of each of these genera and are applying a combination of bioinformatics and experimental approaches to mine the sequences for genes with desirable activities such as enzymes, transporters and secretion systems. We identified a collection of potentially interesting glycosyl hydrolases from these genomes using InterPro and BLAST searches along with information from the CAZy database. In collaboration with Lucigen Corporation we are using high-throughput screening of genomic DNA libraries from these organisms and others to identify clones encoding cellulolytic enzymes. To provide an efficient system for delivering these enzymes from the bacterium into the extracellular milieu we are working on identifying and cloning secretion systems from these bacteria and expressing them in *E. coli*.

A Genome-Wide Analysis of MADS-Box Genes in *Physcomitrella patens*

Elizabeth I. Barker* (barker1e@uregina.ca) and Neil W. Ashton

Biology Department, University of Regina, Saskatchewan, Canada

MADS-box genes perform diverse roles in both reproductive and vegetative plant development but are best known as specifiers of floral meristem and floral organ identities. Furthermore, rapid expansion of the MADS-box gene family is thought to have played an important role in the evolution of angiosperms and the flower that defines them. The presence and roles of floral organ identity *B*- and *C*-function orthologues in non-flowering seed plants indicate that mechanisms governing the differentiation of reproductive structures in all seed plants are fundamentally similar and evolutionarily conserved. It is also intriguing that non-orthologous homologues of these genes have been discovered in a broad range of cryptogams and a role for some of them in sexual reproduction has been demonstrated for the moss, *Physcomitrella patens*, and postulated for three charophycean algae. Revelation of the full complement of MADS-box genes in *Physcomitrella*, facilitated by completion of its genome sequence, has revealed that this gene set is intermediate in size (26 genes) compared to the sets present in the sequenced genomes of green algae (1 gene) and angiosperms (approximately 100 genes). Also, the proportions of $MIKC^C$, MIKC* and Type I genes are strikingly different from those characteristic of angiosperms. Several closely linked pairs or triplets of MADS-box genes in *Physcomitrella* indicate that expansion of this gene family occurred, at least in part, by tandem duplications. Conversely, the arrangement of genes on separate scaffolds and the pattern of phylogenetic clustering of the MADS-box genes indicate that further expansion resulted from segmental duplication events. The exceptionally high degree of conservation of both nucleotide sequence and gene architecture within the $MIKC^C$ and MIKC* subtypes and within each clade of Type I genes suggests that the MADS-box gene complement of *Physcomitrella* was formed mainly by lineage-specific expansion. Phylogenetic analysis of MADS-box genes from a broad range of evolutionarily informative plant taxa supports the postulate that expansion of the MADS-box gene family occurred separately in the moss and seed plant lineages after they diverged. In the latter case, it appears that there was substantial expansion of the MADS-box family both before and after separation of the line to angiosperms. Despite the lack of demonstrable orthology, it remains an open and controversial question whether some members of this gene family in both flowering plants and moss have retained vestiges of an ancestral (reproductive) role of MADS-box genes, which were present in their common (charophycean) progenitor.

Tracking CRISPR Sequences in the Wilderness: Metagenomic Analysis of Acidic Hot Springs in Yellowstone National Park

M.M. Bateson* (mbateson@montana.edu), A.C. Ortmann, V.J.B. Ruigrok, F.F. Roberto, T. Douglas, and M.J. Young

Montana State University, Bozeman

We present a metagenomics approach to assess viral diversity and the cellular defense response to viral interaction. Metagenomes for paired cellular and virus-enriched fractions were obtained from a high temperature acidic hot spring (CHANN041, 2007-April) from Yellowstone National Park (YNP). The low diversity host community is comprised of as few as two dominant archaeal genera while the viral metagenome revealed previously

unknown viruses. Analysis of the cellular metagenome CRISPR-spacer (Clustered Regularly Interspaced Short Palindromic Repeats) sequences associated with the hosts' viral defense matched sequences present in the viral metagenome. A comparison of the CRISPR-spacer sequences from the cellular metagenomes obtained from four other Yellowstone hot springs showed no overlap with the viral metagenome, indicating that each cellular community is directly responding to viruses in their environment. The results support the concept that CRISPRs are actively evolving cellular loci that reflect past and current virus encounters. The study demonstrates that paired cellular-viral metagenomics studies can expand our understanding of virus diversity and the cellular mechanisms contributing to this diversity.

Understanding Nitrogen Limitation in *Aureococcus anophagefferens* Through cDNA and qRT-PCR Analysis

Gry Mine Berg * (mineberg@stanford.edu), Jeff Shrager, Gernot Glockner, Kevin R. Arrigo, and Arthur R. Grossman

Stanford University

Brown tides of the marine pelagophyte *Aureococcus anophagefferens* have been investigated extensively for the past two decades. Its growth is fueled by a variety of nitrogen compounds, with dissolved organic nitrogen being particularly important during blooms. Characterization of nitrogen transporters in a cDNA library suggests that *A. anophagefferens* can assimilate eight different forms of nitrogen. Expression analysis of these transporters demonstrated the highest relative accumulation of a transcript encoding a novel purine transporter followed by the urea active transporter *DUR3*. This suggests that purines, and their degradation product urea, are important sources of nitrogen for the growth of this organism and could possibly contribute to the initiation and maintenance of blooms in the natural environment.

Systematic Approach for the Improvement of Alfalfa Production using Genomic Diversity of *Sinorhizobium meliloti* Natural Populations

Emanuele G. Biondi[1]* (emanuele.biondi@unifi.it), Alessio Mengoni,[1] Matteo Brilli,[2] Marco Bazzicalupo,[1] and Stefano Mocali[3]

[1]Department of Evolutionary Biology, University of Florence, Italy; [2]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, Lyon, France; and [3]C.R.A. – Centro di Ricerca per lo Studio delle Relazioni tra Pianta e Suolo, Via della Navicella, Rome

Alfalfa is a legume crop commonly used as forage or in crop rotation practices providing organic nitrogen to the soil via its symbiosis with the nitrogen fixing bacterium *Sinorhizobium meliloti*. To date the investigation of this agronomic system has been analyzed using classical non-systematic genetic tools. The large phenotypic diversity present in nature, which presents a huge biotechnological interest, has not been genetically explored using powerful post-genomic approaches.

Extensive characterization of 4 natural strains was performed using comparative genomic hybridization (CGH) and phenotype microarray (PM) in order to find genetic determinants

of natural capabilities of those strains but this approach failed. Therefore among those strains analyzed, the efficient nitrogen fixing strain BL225C from cultivated soil of Lodi, Italy, and AK83, a salt resistant isolate from the arid Aral Sea region, are now in the sequencing process at JGI. Results of this genomic project are going to be compared and associated with the phenotypic investigation. Eventually candidate genes of salt resistance and nodulation efficiency will be combined in a single strain in order to create a "super-rhizobium" able to colonize marginal lands but keeping favorable agronomic features.

Better Tools for Interpreting and Presenting Genomes: The GATOR System

Jeffrey L. Boore* (jlboore@genomeprojectsolutions.com) and Susan I. Fuerstenberg

Genome Project Solutions, Hercules, California

We anticipate a great acceleration in whole genome sequencing over the next few years. Current tools for interpreting, comparing, and presenting these data cannot handle the expected pace, lack integration, require extensive IT support and computational expertise, and do too little to facilitate biological discovery. In particular, the standard "browser" format is anachronistic, with the genome assembly, rather than the biological information, being the organizing principle. It requires great manual effort to identify any particular gene or biochemical pathway. Fortunately, new technological developments are now enabling a better approach. First, the cost is now reasonable to generate >1,000,000 ESTs for every genome project, from which a fairly complete gene set can be accurately constructed. Second, we have developed effective tools for assigning orthology among genes based on phylogenetic analysis of all genes. We are building the "GATOR" system (Genome Analysis Tools and Online Resources), a "gene-centric", user-friendly, streamlined approach to genome interpretation, comparison, and presentation. The entry point is the gene catalog itself, sortable by many categories, including domain content, intracellular location, SNP content, biochemical pathway, protein characteristics, or number of members in any gene family. GATOR presents evolutionary trees, gene colinearity maps, and links to protein structures for all genes in multiple sequenced genomes.

Generation of an Insertional Mutant Library in *Brachypodium distachyon*

Jennifer Bragg[1]* (Jennifer.bragg@ars.usda.gov), Jiajie Wu,[2] Yong Gu,[1] Gerard Lazo,[1] Olin Anderson,[1] and John Vogel[1]

[1]USDA ARS, Western Regional Research Center, Albany, California; and [2]University of California, Davis

The objective of our work is to generate >7,500 insertional mutants in the model grass *Brachypodium distachyon* and to sequence the regions flanking >6,000 insertion sites. The location of insertions in the genome will be determined 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. We have performed experiments to optimize transformation, evaluate transposon systems, and compare

transformation efficiencies of vectors built with different promoters, reporter genes, and selectable markers. We have observed considerable variation in efficiency and plant survival and fertility depending on the construct used. Transformations employing hygromycin selection yielded consistently higher efficiency and survival over those using BASTA. For our best constructs, we achieve an average efficiency of 48%, and the mean survival rate is 89% after the plants have been transferred to soil. Transformations using En/Spm transposons resulted in regenerants that died before being moved to soil. Plants generated using Ac/DS transposons had a 33% survival rate in soil, but those that flowered failed to produce viable seed. To date, we have generated >750 T_0 mutant plants. Over 95% of T_0 plants tested showed expression of the GUS reporter gene, and all were positive by PCR for the presence of the selectable marker. T_1 seed has been harvested for 235 plants and will be planted for phenotypic evaluation. Here we describe our procedures and the status of our project.

Chorismate Synthase and Colonization of Xylem Vessels of *B. napus* by the Phytopathogenic Fungus *V. longisporum*

Susanna A. Braus-Stromeyer* (sbraus@gwdg.de), Seema Singh, and Gerhard H. Braus

Institute of Microbiology and Genetics, Georg August University, Goettingen, Germany

Verticillium longisporum is a devastating soil-borne fungal pathogen of oilseed rape (*Brassica napus*). Oilseed rape is one of the most important bio-fuel producers in Europe. The fungus colonizes xylem vessels of host plants. The nutritional status of the fungus in xylem vessels is not yet known. The xylem sap is a water solution containing low concentrations of proteins, sugars and in addition amino acids. The extent of the plant defense mechanisms in the xylem and the subsequent responses of the fungus are largely unexplored.

We analyzed whether biotrophic fungal growth depends on intact aromatic amino acid biosynthesis or is supported by the amino acids provided by the plant xylem. Therefore, we constructed bradytrophic mutants of *V. longisporum* impaired in aromatic amino acid biosynthesis. We knocked down ARO2, the gene for chorismate synthase by RNAi technology. The ARO2 encoded enzyme catalyzes the synthesis of chorismate, the precursor of the three aromatic amino acids. The resulting delta aro2 bradytrophs showed no inhibition during saprophytic growth on minimal medium. In contrast, their virulence in *B. napus* was markedly reduced. Therefore, silenced mutants were able to produce enough chorismate and aromatic amino acids to sustain growth. It is tempting to speculate that additional aromatic compounds deriving from chorismate might be required for the biotrophic life. These might be secondary metabolites required for communication or defense.

In future, we plan to sequence the genome of the *V. longisporum* and analyze the transcriptome and proteome to identify pathogenicity factors of the fungus.

Genome Sequencing of Stalked Bacteria from Aquatic Environments

Pamela J.B. Brown[1]* (pjbonner@indiana.edu), David T. Kysela,[1] Ellen N. Weinzapfel,[1] Sun Kim,[2,3] and Yves V. Brun[1,2]

[1]Department of Biology, [2]Center for Genomics and Bioinformatics, and [3]School of Informatics, Indiana University, Bloomington

Stalked bacteria, such as the model bacterium *Caulobacter crescentus*, synthesize one or more thin extensions of their cell envelope, the prosthecae or stalk, which act as antennae and amplify their ability to take up nutrients from their environment. The narrow stalk adds little volume to the cell, and incoming nutrients are thought to diffuse toward the cell's main body, where nutrients are quickly assimilated by metabolic processes. The goal of this project is to produce a high quality draft sequence of ten genomes of prosthecate bacteria, and two closely related non-prosthecate bacteria, selected to represent an increasingly complex collection of morphologies.

Five draft genomes comprised of fewer than 100 scaffolds have been completed and autoannotated. Four additional draft genomes have been completed; however, additional sequencing is needed to improve the assembly. Phylogenetic analyses using genes from draft genomes confirm that stalks have evolved at least twice in the Alphaproteobacteria. We are currently using experimental approaches to determine if the stalks that evolved independently are structurally or functionally similar. Our phylogenetic analysis also indicates that some closely related stalked and non-stalked bacteria share a common ancestor, suggesting that some bacteria have lost the ability to synthesize stalks or that stalk synthesis depends on specific growth conditions in these species.

We have begun to analyze the draft genomes with respect to mechanisms for the biosynthesis and function of stalks and the extent of conservation of regulatory pathways for stalk biosynthesis. We have found that a phosphorelay system that regulates stalk biogenesis in *Caulobacter crescentus* is conserved in stalked bacteria belonging to the Caulobacterales clade. This system is largely conserved in the non-stalked bacterium, *Brevundimonas diminuta*; however, one gene has a 30 amino acid deletion. We are in the process of determining the significance of this deletion. In addition, we are using our draft genome sequences to explore the link between phosphate starvation and stalk biogenesis in several stalked bacteria.

The long-term goal of this project is to engineer bacteria for use in bioremediation. Since stalks take up diffuse compounds from water sources, this feature could be exploited for bioremediation, specifically the uptake of toxic compounds from contaminated water sources. By engineering the bacteria used in bioremediation to make stalks, their ability to take up pollutants and their uptake efficiency can be improved. Conversely, it should be possible to design specific genera of stalked bacteria to combat contamination in specific environments, either by exploiting their metabolic pathways, or by engineering them with metabolic pathways from other organisms. The knowledge acquired in this project could have additional uses in industry. Bacteria are often used as workhorses in the mass-conversion of one molecule to another. For example, engineering a bacterium to synthesize stalks may improve the speed of uptake of a substrate molecule for conversion, potentially impacting processes such as drug and biofuel production.

The Genome Sequences of the Industrially Relevant Fe(II) Oxidizers *Diaphorobacter* sp. Strain TPSY and *Pseudoglubenkiania* sp. 2002

Kathryne G. Byrne-Bailey* (kbyrne@nature.berkeley.edu), Antinea H. Chair, and John D. Coates

Department of Plant and Microbial Biology, University of California, Berkeley

Ongoing studies indicate the ubiquity and diversity of nitrate-dependent Fe(II)-oxidizing bacteria. During enumeration studies from a variety of sediments, two novel nitrate-dependent Fe(II)-oxidizing Betaproteobacteria, *Diaphorobacter* sp. strain TPSY and *Pseudogulbenkiania* sp. 2002, were isolated and characterized. *Diaphorobacter*/*Acidovorax* species, of the *Comamonaceae* order, are ubiquitous in soil and aqueous environments, and are often found as potential symbionts and pathogens in eukaryotic hosts. *Pseudogulbenkiania* spp. of the order *Neisseriales* are typically eukaryotic pathogens. Both strains grew by nitrate-dependent Fe(II) oxidation mixotrophically with acetate as the carbon source while in addition, strain 2002 also grew lithoatotrophically with Fe(II), CO_2, and nitrate. Both strains also oxidized insoluble U(IV) to soluble U(VI) coupled to nitrate reduction. The genomes of strains 2002 and TPSY were 3.78 Kb (draft) and 3.7 Kb (completed) respectively, with 3573 and 3479 predicted protein coding sequences. At the 16S rRNA level strain TPSY had 99.8 % similarity to *Acidovorax* sp. strain JS42, whereas the closest relative to strain 2002 was *Pseudogulbenkiania subflava* (99.3% similarity).

Preliminary annotation of strain TPSY revealed a Tn21 partial sequence encoding genes known to confer resistance to arsenate and mercury, as well as one complete CRISPR region characterized as belonging to the Nmeni subtype associated with bacteria identified as vertebrate pathogens/ commensals. In contrast, strain 2002 had no identified CRISPR regions but several potentially functional prophage regions. Both bacteria had putative genes such as the Tad operon for biofilm formation and strain 2002 possessed a potential Type VI secretion system for injection of effector molecules into its hosts.

Both strains had approximately 40 cytochrome domain-containing proteins, not differing significantly in numbers from the majority of known iron oxidizers/reducers. Strain 2002 appears to have a more extensive suite of cytochromes, including B5 and a Ni-Fe hydrogenase b-type cytochrome (not observed in strain TPSY) which may have a role in U(IV) oxidation.

Although widespread, the microorganisms responsible for Fe(II) oxidation in these environmental systems are virtually unknown, as are the biochemical and genetic mechanism(s) involved. The genetics of anaerobic Fe(II) oxidation are important on a global scale. Recent evidence indicated that anaerobic Fe(II) oxidation can contribute to a dynamic anoxic iron redox cycle affecting soil and sediment biogeochemistry, mineralogy, and heavy-metal and radionuclide immobilization. Therefore, it is hoped the genetic information contained in these genome sequences can be utilized to further investigate iron biogeochemical cycling in natural environments.

Designing Synthetic Riboregulators to Program Gene Expression in Engineered Metabolic Pathways

James M. Carothers* (jmcarothers@lbl.gov), Yuvraaj Kapoor, Jonathan A. Goler, Yisheng Kang, and Jay D. Keasling

California Institute for Quantitative Biosciences and Berkeley Center for Synthetic Biology, University of California, Berkeley, and Joint BioEnergy Institute, Emeryville, California

Developing the ability to quantitatively model and design genetic controls for synthetic biological systems will greatly increase the speed and efficacy with which cellular metabolism can be harnessed to produce industrially-important small molecules. Although a variety of synthetic RNA-based genetic controls (riboregulators) have been created, it has not yet been possible to rationally design their functions in specific biological contexts. Here, we demonstrate that mechanistic models can be used to search biochemical parameter space to guide the assembly of riboregulatory systems that meet specific performance criteria. We show that riboregulated genetic controls employing self-cleaving catalytic RNA structures (ribozymes and aptazymes) to modulate mRNA degradation can be engineered such that their function in *E. coli* can be quantitatively simulated and predicted. In on-going work we are using riboregulated controls to systematically vary enzyme levels in a pathway engineered to produce the precursors of isoprenoid-derived biofuels. Quantitatively mapping the relationships between mRNA and protein levels, enzyme kinetics and fluxes will help optimize production in this pathway and functionally characterize system components for use in future applications.

Community-Involved Microbial Genome Analysis at JGI-LANL

Jean Challacombe* (jchalla@lanl.gov), Gary Xie, Monica Misra, Susana Delano, Thomas Brettin, and Chris Detter

Los Alamos National Laboratory, Los Alamos, New Mexico

The role of the genome analysis team at JGI-LANL is to facilitate publication of JGI genome papers and provide bioinformatics support and training to promote community-involved genome analysis. Our team members work closely with JGI collaborators on the comparative analysis of their genomes, with the goal of publishing the results in high profile scientific journals. When a JGI collaborator needs help with analysis and paper preparation, we assign an analysis team member to their project, in either a leading or supporting role. If we are leading the effort, we design and do most of the analysis, write the paper, and take the first author position on the paper. If we support the collaborator's effort, we perform specialized analyses, but someone from the collaborator's lab takes the lead on the paper. We also offer bioinformatics training for those collaborators who want to take the lead on the analysis and paper writing effort. Since 2005, we have conducted on-site, week-long, bioinformatics training sessions for 10 JGI collaborators, students and postdocs. Over the past 4 years, we have hosted visits by 36 JGI collaborators as part of our Genomic Explorers Seminar Series. We routinely receive requests for specialized analyses from JGI collaborator labs. In projects where JGI-LANL team members played a leading role in the analysis and preparation of genome papers, 10 genome papers have been published and 4 manuscripts are in preparation.

Transcriptional Profiling of *Plantago ovata* to Discover Genes Involved in Arabinoxylan Biosynthesis

Jean-Christophe Cocuron[1] and **Curtis G. Wilkerson**[1,2]* (wilker13@msu.edu)

[1]Department of Plant Biology and [2]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing

Arabinoxylan is a major cell wall component in grasses. As an example, destarched corn fiber contains approximately 70% arabinoxylan (Doner and Hicks 1997). The use of grass species to produce biomass for ethanol production is complicated by the large amounts of pentoses that result from the conversion of arabinoxylan to arabinose and xylose. Current fermentation systems do not effectively ferment these sugars. Alterations in the amount of arabinoxylan in plants used for ethanol production could produce substantial benefits in the conversion of biomass to ethanol. In order to achieve alterations in arabinoxylan content we need have a better understanding of arabinoxylan biosynthesis than currently exists. One way to discover enzymes and regulatory proteins involved in a biological pathway is to examine the mRNAs of cells actively engage in the pathway of interest compared to mRNAs isolated from cells not engage in the pathway. Such transcriptional profiling is most successful when the pathway of interest is a major activity of the tissue. The mucilagous layer of psyllium (*Plantago ovata* Forsk) seed contains about 60% arabinoxylan by weight (Fischer et al. 2004) and so represents a good tissue to use for transcriptional profiling to discover genes involved in arabinoxylan biosynthesis. We have successfully generated 4 cDNA libraries 6, 8, 10 and 12 days post anthesis (DPA) and have obtained over 850,000 sequences using the Roche GS-FLX sequencer. An examination of these sequences revealed that enzymes involved in the biosynthesis of UDP-xylose were highly represented in those cDNA libraries suggesting that these libraries likely are enriched in genes involved in arabinoxylan biosynthesis and its regulation. We are currently expressing candidate genes in heterologous systems and assaying the expressed proteins for xylan synthase and arabinosyltransferase activity. We are also examining *Arabidopsis* knockout mutants for changes in cell wall composition.

Directed Evolution of Ionizing Radiation Resistance in *Escherichia coli*

Michael M. Cox[1]* (cox@biochem.wisc.edu), Dennis R. Harris,[1] Steve V. Pollock,[2] Elizabeth A. Wood,[1] Reece J. Goiffon,[1] Audrey J. Klingele,[1] Julie Eggington,[1] Trang D. Nguyen,[2] Christina M. Middle,[3] Jason E. Norton,[3] Eric L. Cabot,[4] Michael C. Popelars,[1] Hao Li,[1] Sarit A. Klugman,[1] Lindsay L. Hamilton,[1] Lukas B. Bane,[1] Cassandra Jenson,[1] Joel Martin,[5] Wendy Schackwitz,[5] Len Pennacchio,[5] Nicole T. Perna,[4,6] Thomas J. Albert,[3] and John R. Battista[2]

[1]Department of Biochemistry, University of Wisconsin, Madison; [2]Department of Biological Sciences, Louisiana State University and A & M College, Baton Rouge; [3]Roche NimbleGen Inc, Madison, Wisconsin; [4]Genome Center, University of Wisconsin, Madison; [5]DOE Joint Genome Institute, Walnut Creek, California; and [6]Laboratory of Genetics, University of Wisconsin, Madison

Four extremely radioresistant populations of *Escherichia coli* K12 were independently derived from MG1655 following repeated exposure to high dose ionizing radiation. D_{37} values for strains isolated from two of the populations approached that exhibited by *Deinococcus radiodurans*. Complete genome re-sequencing using two different methods

revealed that strains taken from independently evolved populations acquired radioresistance through very different sets of genetic changes. The results do not reinforce any current idea about the mechanisms underlying the radiation resistance of species such as *Deinococcus radiodurans*. Complete genome resequencing of multiple strains isolated from one evolved population revealed extensive mutational diversity, but also common patterns that imply mechanisms underlying the new phenotype. In this population, genomic changes are concentrated in genes involved in recombinational DNA repair, replication restart, glutamate transport, and regulation of cell division. There appear to be multiple paths to radiation resistance, and multiple contributing mechanisms. Two significant positive contributions to the acquired phenotype are documented, (a) deletion of prophage e14 (the only genome alteration that is common to all of the evolved strains), and (b) certain mutations in the *recA* gene (D276A and A289S). Mutational epistasis is also evident. A mutation in the *mutT* gene, common enough in an evolved population to indicate positive selection, is deleterious to the IR resistance phenotype when isolated in an otherwise wild type background.

The Unicellular Green Alga *Dunaliella salina* as a Model System for Studies of Stress Metabolism

John Cushman[1] and **Juergen Polle**[2]* (JPolle@brooklyn.cuny.edu)

[1]Department of Biochemistry and Molecular Biology, University of Nevada, Reno; and [2]Department of Biology, Brooklyn College of City University of New York

Unicellular green algae of the genus *Dunaliella* are divided into two subgenera: *Pascheria* and *Dunaliella* (Masjuk 1972; Preisig 1992). The subgenus *Dunaliella* contains species that are remarkably halotolerant, i.e. their growth range includes NaCl concentrations that are lower than seawater <0.5 M or reach saturation levels with >5 M. Consequently, *Dunaliella* algae were used for a long time to study how cells cope with a wide range of salinities. Other research areas included for example over-accumulation of secondary carotenoids in the plastid of *Dunaliella salina* or regulation of photosynthesis in response to environmental stress. Most recently unicellular green algae of the genus *Dunaliella* have received renewed attention, specifically in the area of their potential for biofuels production. An overview of the current status of development of *D. salina* as a model system will be presented.

Transcriptome Analysis of *Chlamydomonas reinhardtii* Using Ultra-High-Throughput Sequencing

David Casero Díaz-Cano,[1] Madeli Castruita,[2] **Steven Karpowicz**[2]* (skarp@chem.ucla.edu), Sabeeha Merchant,[2] and Matteo Pellegrini[1]

[1]Department of Molecular, Cell and Developmental Biology and [2]Department of Chemistry and Biochemistry, University of California, Los Angeles

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. 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. We evaluate the potential of Illumina's Solexa technology for generating a whole transcriptome for *Chlamydomonas* and use the resulting data as the basis for gene model prediction.

Sample Shipping and Storage at JGI

Erin Dunwell* (emdunwell@lbl.gov), Meric Velasco, Justin Hatch, and Tijana Glavina del Rio

DOE Joint Genome Institute, Walnut Creek, California

JGI's Freezer Farm group is responsible for storage and shipping of samples post sequencing. This poster will include variety of information regarding sample shipping such as names of institutions and countries we ship to and rate of repetition, cost associated with shipping, time line for shipments from onset to destination, number of plates shipped out per year. In addition to shipping, the freezer farm is also responsible for storage and plate management of samples post sequencing. The PMO office and Freezer Farm group work closely in managing the sample inventory in order to maintain a manageable amount of samples at JGI. We will include information regarding the number of on site and off site freezers we use for sample storage and how the two facilities are managed. This type of work is highly repetitive so it is important to keep our staff safe and healthy. Therefore a small part of this poster will be dedicated to safety and ergonomics specific to the freezer farm.

Evolution of the Mating Type Chromosome in *Neurospora tetrasperma*

Christopher E. Ellison[1]* (cellison@berkeley.edu), Jason E. Stajich,[1] David Jacobson,[1] Alla Lapidus,[2] Brian Foster,[2] and John W. Taylor[1]

[1]Department of Plant and Microbial Biology, University of California, Berkeley; and [2]Microbial Genomics DOE Joint Genome Institute, Walnut Creek, California

Neurospora tetrasperma is unique among species of *Neurospora* in that, after meiosis, two haploid nuclei of opposite mating type are packaged into a single ascospore. This pseudohomothallic condition results in self-fertility. Blocked recombination proximal to the mating type locus during meiosis I is necessary for the correct packaging of alternative mating type alleles into a single heterokaryotic ascospore. Previous genetic studies have shown that recombination is suppressed along the majority of the mating type chromosome (75% / 7Mbp) possibly due to rearrangements within the *N. tetrasperma mat A* chromosome. The Joint Genome Institute has recently sequenced both mating types of *N. tetrasperma* (*mat a* [FGSC #2509] and *mat A* [FGSC #2508]) at 0.5X and 8X respective coverage. In addition, two runs of 454 pyrosequencing (one standard and one paired-ended) were completed for *N. tetrasperma mat A*, increasing coverage to 23X and dramatically improving the assembly. Here we show that a series of intrachromosomal rearrangements have occurred on the *N. tetrasperma mat A* chromosome with respect to

N. crassa mat A, while the autosomes remain essentially collinear. We suggest that these rearrangements may be contributing to the *mat* chromosome recombination block and we plan to investigate the characteristics of the identified breakpoints.

Community Diversity Estimates are Not Affected by Primer Pair or Amplicon Length

Anna Engelbrektson[1]* (aengelbrektson@lbl.gov), Victor Kunin,[1] Natasha Zvenigorodsky,[1] Feng Chen,[1] Howard Ochman,[2] and Philip Hugenholtz[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; and [2]Department of Biochemistry, University of Arizona, Tuscon

16S rDNA-based analyses of microbial communities have historically involved Sanger sequencing. Recent advances have led to the emergence of pyrosequencing as the method of choice in that this technology can, for similar costs, yield at least two orders of magnitude more reads and employ barcoding so that multiple samples are multiplexed in a single sequencing run. Given the depth of sequencing afforded by pyroseqeuncing methods, it is now possible to evaluate several technical aspects of community diversity studies that were previously not possible to investigate. For example, it has been speculated that shorter amplicon lengths may increase diversity estimates due to PCR reaction kinetics. In addition, there is anecdotal evidence that the particular PCR-primer pair will influence diversity estimates. Using the relatively simple microbial community residing in the termite hindgut, we investigated the influence of both primer pair and amplicon length on estimates of species abundance and distribution. We targeted two regions of the 16S rDNA molecule, each using one of two common primers (27 forward or 1392 reverse) paired with complementary primers that yielded amplicons ranging from 278 to 880 bp. The resulting amplicons were sequenced outwards from the common primer by 454 FLX pyrosequencing technology. Surprisingly, we found that neither amplicon length and primer pair within a region had an appreciable effect on community composition metrics. The largest effect on diversity estimates was observed between the two targeted portions 16S attributable to broad differences in the level of sequence variation within each region. These results highlight the need to examine common 16S regions when comparing diversity estimates across communities, but also indicate that different-sized amplicons spanning a common region can be accurately compared across samples or environments.

Community Genomic and Proteomic Analysis of Chemoautotrophic, Iron-Oxidizing "*Leptospirillum rubarum*" (Group II) and *Leptospirillum ferrodiazotrophum* (Group III) in Acid Mine Drainage Biofilms

Daniela S. Aliaga Goltsman* (dgolts@eps.berkeley.edu), Vincent J. Denef, Steven W. Singer, Nathan C. VerBerkmoes, Mark Lefsrud, Ryan Mueller, Gregory J. Dick, Christine Sun, Korin Wheeler, Adam Zemla, Brett J. Baker, Loren Hauser, Miriam Land, Manesh B. Shah, Michael P. Thelen, Robert L. Hettich, and Jillian F. Banfield

University of California, Berkeley

We analyzed near-complete population (composite) genomic sequences for coexisting acidophilic iron-oxidizing *Leptospirillum* Groups II and III bacteria (phylum Nitrospirae) and an extrachromosomal plasmid from a Richmond Mine, CA acid mine drainage (AMD) biofilm. Community proteomic analysis of the genomically characterized sample and two other biofilms identified 64.6% and 44.9% of the predicted proteins of *Leptospirillum* Groups II and III, respectively and 20% of the predicted plasmid proteins. The bacteria share 92% 16S rRNA gene sequence identity and > 60% of their genes, including integrated plasmid-like regions. The extrachromosomal plasmid encodes conjugation genes with detectable sequence similarity to genes in the integrated conjugative plasmid, but only those on the extrachromosomal element were identified by proteomics. Both have genes for community-essential functions, including carbon fixation, biosynthesis of vitamins, fatty acids and biopolymers (including cellulose); proteomic analyses reveal these activities. Both *Leptospirillum* types have multiple pathways for osmotic protection. Although both are motile, signal transduction and methyl-accepting chemotaxis proteins are more abundant in *Leptospirillum* Group III, consistent with its distribution in gradients within biofilms. Interestingly, *Leptospirillum* Group II uses a methyl-dependent and *Leptospirillum* Group III a methyl-independent response pathway. Although only *Leptospirillum* Group III can fix nitrogen, these proteins were not identified by proteomics. Abundances of core proteins are similar in all communities, but abundance levels of unique and shared proteins of unknown function vary. Some proteins unique to one organism were highly expressed and may be key to the functional and ecological differentiation of *Leptospirillum* Groups II and III.

Genomics Research and Undergraduate Education–A Value-Added Collaboration in Both Directions

Stuart Gordon* (gordonsg@hiram.edu), Kathryn Reynolds, and Brad Goodner

Hiram College, Hiram, Ohio

Integrating original research into a course is an effective way to connect students to the current state of understanding, encourage them to take more control of their own learning, and promote problem solving and interdisciplinary learning. Over the past 8 years at Hiram College, we have integrated genome annotation, functional genomics and metagenomics into Molecular and Cellular Biology, Genetics, and Microbiology courses. Working in teams, students have complemented the usual automated first-pass annotation (gene localization and identification through similarity) by reconstructing biochemical pathways to verify gene-protein relationships and to identify novelties and redundancies, by predicting regulatory networks based on gene order and shared non-coding sequences, by identifying possible instances of lateral gene transfer based on gene phylogenies, and by comparing large gene sets across genomes to test various hypotheses. We will highlight the work done by the 2007-2009 iterations of these courses which include our participation in the JGI IMG-ACT pilot program.

It is also important for students to see genome annotation not only as answers to some past questions but also as generators of new testable hypotheses which can often be addressed by undergraduates and even high school students. For example, students in the Molecular and Cellular Biology course have used reverse genetics to test functional predictions based on bioinformatics analyses of over 75 genes in *Agrobacterium tumefaciens*. Students in the Genetics course, along with high school students during the past several years, have used forward genetics to link hundreds of genes to functions through large scale transposon mutant hunts in *A. tumefaciens*, *Chromohalobacter salexigens*, and *Acidovorax avenae*

subsp *avenae*. More recently, Genetics students have used functional complementation of known *E. coli* mutants to find functional homologs from known genomes and from metagenomes. Finally, summer research students have used comparative genomic data to design PCR primers targeted for different phylogenetic groups as part of an ecological metagenomics project. These primers were successfully tested in the Genetics course and will contribute to future Microbiology courses and high school outreach projects. We will highlight some of the more interesting findings in several functional categories.

Seasonal Dynamics of Antarctic Bacterioplankton Revealed by Metagenomics

J.J. Grzymski* (joeg@dri.edu), C.S. Riesenfeld, and A.E. Murray

Earth and Ecosystem Sciences, Desert Research Institute, Reno, Nevada

Background: The Antarctic Peninsula marine waters (~64°S latitude) experience strong temporal variability in physical parameters which drive strong gradients in biological productivity in the upper ocean. Little is known about the bacterial and archaeal metabolic and genomic differences between the period of maximal productivity in the summer and the Antarctic winter. **Methods**: Two large-insert metagenomic libraries were created from marine bacterioplankton (<2.5 μm fraction) from February (summer) and August (winter) samples collected in Antarctic Peninsula waters. Bi-directional end-sequencing (~37,000 reads) generated 7 and 9.6 Mb of coding DNA for the summer and winter libraries, respectively. **Results**: A sizeable fraction of the data represents genetic diversity not currently represented in large public databases, while other sequences resemble portions of completely sequenced marine archaeal and bacterial genomes (e.g., *Nitrosopumilus maritimus*, *Polaribacter irgensii*, *Roseobacter denitrificans* and *Pelagibacter ubique*). Comparison of the two metagenomic libraries revealed differences in overall nucleotide composition, phylogenetic distribution of BLAST hits and predicted functions (e.g., frequency of protein families). One conspicuous difference between the libraries was the GC% distribution of end-sequence reads. In the case of the summer sample, the distribution was skewed towards GC-rich sequences, which can be explained (in part) by the fact that *Roseobacter*-affiliated sequences were abundant in the summer sample. Sequence clusters were identified that affiliated exclusively with the summer (360 COGs) or winter (584 COGs) data sets. The Antarctic winter library encodes a more diversified metabolic capacity than the summer library, including pathways for carbon fixation, and nitrogen and sulfur utilization that were not detected in the summer library whereas the summer metagenome encodes more signal transduction and transport capabilities as well as a suite of carbon metabolism and organic phosphorus utilization pathways. **Conclusion**: The Antarctic summer and winter metagenomes revealed major differences in diversity of organisms and functions in temporally distinct marine bacterioplankton communities.

CAM Plants for Biofuel from Marginal Lands: Progress Towards Understanding the Functional Genomics of CAM in High-Productivity CAM Species

James Hartwell[1]* (hartwell@liv.ac.uk), Anne Borland,[2] Howard Griffiths,[3] and Andrew Smith[4]

[1]School of Biological Sciences, University of Liverpool, United Kingdom; [2]Institute for Research on Environment and Sustainability, University of Newcastle, United Kingdom; [3]Department of Plant Sciences, University of Cambridge, United Kingdom; and [4]Department of Plant Sciences, University of Oxford, United Kingdom

The metabolic adaptation of photosynthesis known as Crassulacean acid metabolism (CAM) permits the net uptake of CO_2 at night and the highly efficient use of available water. This photosynthetic pathway is expressed in ~7% of higher plant species, many of which dominate the biomass in arid and semi-arid regions of the world. Cultivated CAM species including *Opuntia ficus-indica, Agave tequilana,* and pineapple can achieve high productivity in regions where limited precipitation and high evapotranspiration rates preclude the cultivation of C_3 and C_4 crops.

The potential of *Agave* as an economically viable source of remarkably high yields of bioethanol has recently been reported from Mexico. This poster highlights the key attributes of CAM that contribute towards high biomass and bioethanol production from marginal lands. Areas of current and future research are outlined, including progress on the functional and comparative genomics of CAM. In particular, an update will be presented on the *Kalanchoe fedtschenkoi* transcriptome sequencing project, which is using Roche/454 GS-FLX Titanium and Applied Biosystems SOLiD next generation sequencers to identify the genes required for CAM in *K. fedtschenkoi*. This project will ultimately provide a systems level view of CAM that will inform and improve the potential of CAM plants for carbon sequestration and biofuel production on marginal lands.

The *K. fedtschenkoi* transcriptome sequencing project in the JH lab is supported by BBSRC grant BB/F009313/1

Fosmid Library Construction for Switchgrass (*Panicum virgatum*): A Resource for Analysis of Genes Governing Bioenergy-Related Traits

Jennifer S. Hawkins* (jhawkins@uga.edu), Ryan J. Percifield, Matt Estep, and Jeffrey L. Bennetzen

Departments of Genetics and Plant Biology, University of Georgia, Athens

Panicum virgatum, commonly known as switchgrass, has been identified as an attractive biomass source for the production of ethanol from lignocellulosic materials. We have constructed a 6x coverage fosmid library for switchgrass to facilitate sequence analysis of genes involved in bioenergy-related traits. The fosmid library is comprised of almost 200,000 clones, each containing a 35-40 kb fragment of the switchgrass genome, for a total of ~7 billion base pairs of switchgrass DNA. The fosmids were pooled in such a way as to simplify the identification process of clones containing genes of interest through three PCR amplification steps. To determine the quality and utility of the library, and to isolate a potentially useful set of promoters for transgene expression, degenerate primers

homologous to *ubiquitin* were used to screen and identify clones of interest. A *ubiquitin*-containing fosmid was sequenced and assembled into a single full-length contig of 37,820 bp that contains two tandem *ubiquitin* genes, three other putative genes and two retrotransposons.

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

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

Feedstocks Division, Joint BioEnergy Institute, Emeryville, California

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 assess 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.

Discovery of Substrate-Targeted Enzymes for the Degradation of Biomass by Metatranscriptomics

M. Hess[1]* (mhess@lbl.gov), T. Zhang,[1] S.G. Tringe,[1] R. Mackie,[2] and E.M. Rubin[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; and [2]University of Illinois, Urbana-Champaign

Cellulolytic enzymes that are highly active and stable represent major bottlenecks for the efficient large-scale production of biofuels from lignocellulose. The bovine rumen is a complex microbial habitat known to harbor fibrolytic microbes and represents a promising source of novel biocatalysts for lignocellulose degradation. We employed high-throughput pyrosequencing to identify feedstock-targeted enzymes within the transcriptome of bovine rumen microbial communities.

Switchgrass and alfalfa were incubated for 3 days in the bovine rumen. Nucleic acids were extracted from the microbial communities adhered to switchgrass and alfalfa to study whether specific microbes were associated with particular substrates. Results obtained by 16S ribosomal RNA sequencing showed that the microbial community tightly associated with switchgrass is significantly different from the community associated with alfalfa. This

suggests that a distinct set of organisms is involved in the degradation of each of these two feedstocks.

Expression profiling of the switchgrass associated microbiome was performed by massively parallel sequencing and almost 300 putative glycosyl hydrolases (GHs) were identified. Some of these GHs might be highly specialized for the degradation of switchgrass. We will soon analyze the expression profiles of microorganisms adherent to other biofuel crops (i.e. miscanthus and corn stover), thereby identifying microbial genes and the corresponding proteins that might contribute to the efficient degradation of these substrates.

The results obtained in the course of our study indicate that the fiber-bound microbes are indeed a rich source of putative cellulolytic enzymes. Heterologous gene expression of the identified GHs genes and detailed physicochemical characterization of the recombinant gene products will allow us to verify the sequence-based annotation of the transcript tags.

Improvements and New Applications for the Illumina Sequencing System

David W. Hillman* (dwhillman@lbl.gov), Mary Ann Pedraza, Sirisha Sunkara, Maria Shin, Jeff Froula, Feng Chen, and Len Pennacchio

DOE Joint Genome Institute, Walnut Creek, California

Next Generation sequencing products and procedures for the Illumina platform were evaluated, several of which immediately benefited JGI users. Paired end-reads of 35 bases each have become routine on our new GAii instruments with output in excess of 5 billion bases per run. Single read genomic sequence data has been used for finishing and polishing for over 2 years, and collaborative work has progressed to include genomic SNP detection, expression analysis for eukaryotes and prokaryotes, and ChIP DNA sequencing. Paired-End reads have been particularly useful for SNP detection. Our beta testing activities include positive results for barcode libraries (indexing) using Illumina's 3 primer scheme, large-gap libraries produced several ways, and accurate 75 base read-length. Key technologies include improvements to the instruments, reagents, software, and information management tools.

Linking Undergraduate Research to the Discovery of Novel Diazotrophs and PGPRs

A.M. Hirsch* (ahirsch@ucla.edu), A.R. Schwartz, E. McDonald, E.R. Sanders-Lorenz, and the students of MCDB120L

Department of Molecular, Cell and Developmental Biology, Molecular Biology Institute, and Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles

Undergraduate science courses that concentrate solely on lecture material or “cook book” laboratories do not mirror the process of scientific research in the 21st century, which is hypothesis-driven, team-building, and discovery-based. We launched two project-based laboratory courses for undergraduates at UCLA to give the students more experience in the process of science: MIMG (Microbiology, Immunology and Molecular Genetics) 121A,

where students monitor the rhizosphere for nitrogen-fixing bacteria, and MCDB (Molecular, Cell and Developmental Biology) 120L, a plant biology course, where the students test the bacteria from MIMG 121A for their ability to promote plant growth either by nitrogen fixation or by secretion of growth factors. So far, several nitrogen-fixing bacteria and PGPRs (Plant Growth Promoting Rhizobacteria) have been discovered from the combined class efforts. In addition to rhizobia, several *Streptomyces* strains were found to grow on –N medium and test positively on a western blot probed with a nitrogenase antibody. The best candidate for further study was *Bacillus simplex*, which has not been identified previously as a PGPR. When the model legumes *Medicago truncatula* and *Lotus japonicus* were inoculated with *B. simplex*, a significant increase in plant biomass was measured. Whether the increase in biomass is a consequence of nitrogen fixation or the synthesis of plant-required growth factors is not known at this time. Although the bacteria grow on a –N medium and degenerate *nifH* primers amplify a PCR fragment of the correct size, sequence analysis shows no relationship of the excised band to *nifH*. Studies on the secretion of plant growth factors such as the plant hormones indole acetic acid (IAA) and cytokinin are in progress. Like MIMG121A, MCDB120 students also use bioinformatics techniques (IMG/ACT) to annotate microbial genomes. By analyzing the multitude of databases presented through IMG/ACT, the students have come to appreciate the many facets of information that one protein can hold, and have discovered how they, as undergraduate students, can contribute to classifying and categorizing genes/proteins.

Annotation of Translation Initiation Sites Using Prodigal

Doug Hyatt, Miriam Land, and **Loren Hauser*** (hauserlj@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

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 examining 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/).

The YNP Metagenome Project: Random Shotgun Sequencing of High-Temperature Chemotrophic and Phototrophic Microbial Communities in Yellowstone National Park

W. Inskeep[1]* (binskeep@montana.edu), Z. Jay,[1] S. Tringe,[2] K. Berry,[2] S. Boomer,[3] D. Bryant,[4] B. Fouke,[5] N. Hamamura,[6] C. Klatt,[1] R. Macur,[1] T. McDermott,[1] D. Mead,[7] S. Miller,[8] N. Parenteau,[9] A-L. Reysenbach,[6] F. Roberto,[10] J. Spear,[11] C. Takacs-Vesbach,[12] D. Ward,[1] and M. Young[1]

[1]Montana State University, Bozeman; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Western Oregon University, Monmouth; [4]Pennsylvania State University, University Park; [5]University of Illinois, Urbana; [6]Portland State University, Portland; [7]Lucigen Corporation, Middleton, Wisconsin; [8]University of Montana, Missoula; [9]NASA Ames Research Center, Mountain View, California; [10]Idaho National Laboratory, Idaho Falls; [11]Colorado School of Mines, Golden, Colorado; and [12]University of New Mexico, Albuquerque

The Yellowstone caldera contains the most numerous and diverse geothermal systems on Earth, yielding an extensive array of unique high-temperature environments that host numerous deeply-rooted and understudied *Bacteria*, *Archaea* and viruses. The *Yellowstone Metagenome Project* is a collaborative effort that was catalyzed in part through networking activities of the NSF Research Coordination Network focused on *Geothermal Biology and Geochemistry in Yellowstone National Park (YNP)*. The primary goal of this DOE-JGI Community Sequencing Project is to acquire and analyze a comprehensive metagenomic dataset of the diverse thermophilic prokaryotic communities inhabiting geochemically distinct geothermal sites within the Yellowstone geothermal complex. Twenty geothermal sites were selected to achieve a representative range of geochemical and physical conditions common in YNP (www.rcn.montana.edu), and to capitalize on an extensive foundation of prior (and on-going) research and characterization of these low-diversity microbial systems.

Twenty geothermal microbial mats and or sediments were sampled in 2007 and 2008 and used to extract DNA of sufficient quantity and quality for construction of small insert libraries, which were subjected to Sanger sequencing at DOE-JGI. Approximately 35-50 megabases of sequence were generated for each site. Aqueous and solid phase samples were also collected for analysis of predominant geochemical constituents and mineralogy associated with each microbial community. Preliminary analysis of random shotgun sequence data from chemotrophic environments (65-88 C) reveal dominant archaeal populations within the Crenarchaeota (e.g, orders Sulfolobales, Thermoproteales, Desulfurococcales, Cenarchaeales and other currently uncharacterized groups) and Euryarchaeota (distantly related to members of the Thermoplasmatales). The predominant bacteria within the chemotrophic habitats included in this study are members of the order Aquificales, and significant diversity within this group exists both within and across different geothermal systems. Other less-dominant bacterial species noted in high-temperature chemotrophic systems include members of the Proteobacteria, Firmicutes, Deinococcus-Thermus, Thermotogae, and other uncharacterized groups. The diverse phototrophic systems represented in the study include both oxygenic and anoxygenic mats ranging in temperature from 48-66 C, where members of the Chloroflexi and Cyanobacteria are common community members across highly diverse environments.

The finished metagenomic data generated for these twenty environments will provide an excellent foundation for understanding microbial diversity and function in extreme thermophilic habitats, as well as potential applications in bio-energy and bio-processing. The YNP Working Group met just preceding the 2009 JGI User's Meeting to discuss data analysis approaches, data interpretation, IMG utilities, and to outline several manuscripts intended to synthesize phylogenetic and functional interpretations of indigenous genomic content across different thermophilic habitats in YNP.

Transcriptional Regulation of Plant Cell Wall Polysaccharide Biosynthesis

Jacob K. Jensen[1]* (jensen58@msu.edu), Jean-Christophe Cocuron,[1] Yan Wang,[1] Nick Thrower,[1] Linda Danhof,[1] Cliff Foster,[1] Curtis G. Wilkerson,[1,2,3] Kenneth Keegstra,[1,2,3,4] and Markus Pauly[1,2,4]

[1]Great Lakes Bioenergy Research Center, [2]MSU-DOE Plant Research Lab, [3]Department of Plant Biology, and [4]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing

Certain seeds produce massive amounts of a single polysaccharide also found in the cell walls of plants as a storage polymer during their development. Deep sequencing of the transcriptome in those seeds should allow the identification of the entire biosynthetic pathway of the particular polysaccharide and its regulation. Here we are investigating Nasturtium (*Tropaeolum majus*) seeds that produce the hemicellulose xyloglucan.

The deposition of xyloglucan in Nasturtium seeds occurs over 9 days late in development. To analyze of the transcriptome in this period of time we are first taking advantage of the 454 Life Sciences Genome Sequencer FLX to sequence total cDNA populations and obtain sequence contigs representing the transcriptome in the seed. We then subsequently use Solexa RNAseq sequencing (Illumina Genome Analyzer II) that gives higher EST counts but shorter sequence reads to obtain high resolution expression profiles from multiple stages of the seed during xyloglucan deposition.

At present we have obtained and assembled 1.6 million ESTs generated with 454 technology covering both early and late stages of xyloglucan deposition. We expect to reach a total of 2 million ESTs. The reads have an average length of 220 bp and give rise to 30,000 assembled contigs. Solexa RNAseq sequencing resulted in a total of 52 million ESTs and were preformed on 7 discrete stages during xyloglucan deposition. We are currently in the process of pasting the 36 bp Solexa reads over the 454 contig backbone.

Preliminary analysis of the sequencing data has lead to essential insights into xyloglucan biosynthesis including glycansynthases and glycosyltransferases, precursor pathways and a handful of candidate transcription factors that might be involved in regulation the transcription of the entire xyloglucan biosynthetic machinery.

Construction of a Consolidated Bioprocessor Derived from *Escherichia coli*

David H. Keating,* Mary Tremaine, and Robert Landick

Great Lakes Bioenergy Research Center, University of Wisconsin, Madison

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.

The Plastic Finished Genome of the Fungal Wheat Pathogen *Mycosphaerella graminicola*

Gerrit H.J. Kema[1]* (gert.kema@wur.nl), Stephen B. Goodwin,[2] Sarrah Ben M'Barek,[1,3] Theo A.J. Van der Lee,[1] and Alexander H.J. Wittenberg[1,4]

[1]Plant Research International B.V., Wageningen, The Netherlands; [2]USDA ARS, Crop Production and Pest Control Research Unit and Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana; [3]Graduate School Experimental Plant Sciences, Wageningen, The Netherlands; and [4]Wageningen University, Department of Plant Breeding, The Netherlands

Meiosis in the haploid plant-pathogenic fungus *Mycosphaerella graminicola* results in eight ascospores due to a mitotic division following the two meiotic divisions. The transient diploid phase allows for recombination among homologous chromosomes. However, some chromosomes of *M. graminicola* lack homologs and do not pair during meiosis. Because these chromosomes are not present universally in the genome of the organism they can be considered to be dispensable. Detailed genetic analyses of two high density mapping populations revealed that *M. graminicola* has 21 chromosomes including up to eight dispensable chromosomes, the highest number reported in filamentous fungi. These chromosomes vary from 0.41 to 0.77 Mb in size, representing 38% of the chromosome number and 11.6% of the genome. Chromosome numbers among progeny isolates varied widely, with some progeny missing up to three chromosomes, while other strains were disomic for one or more chromosomes. Between 15-20% of the progeny isolates lacked one or more chromosomes that were present in both parents. The two high-density maps showed no recombination of dispensable chromosomes and hence, their meiotic processing may require distributive disjunction, a phenomenon that is rarely observed in fungi. The maps also enabled the identification of individual twin isolates from a single ascus that shared the same missing or doubled chromosomes indicating that the chromosomal polymorphisms were mitotically stable and originated from nondisjunction during the second division and, less frequently, during the first division of fungal meiosis. High genome plasticity could be among the strategies enabling this versatile pathogen to quickly overcome adverse biotic and abiotic conditions in wheat fields. Additionally, we used a Comparative Genomic Hybridization whole-genome array based on the finished genome of *M. graminicola* (http://genome.jgi-psf.org). This confirmed that chromosomes 14-21 were frequently absent among isolates, without visible effect on viability or virulence, whereas chromosomes 1-13 were invariably present. The dispensable chromosomes are smaller and have significantly lower gene densities. Most of their genes are duplicated on the essential chromosomes and show a different codon usage. Dispensable chromosomes also contained a higher density of transposons, pseudogenes, and unclassified genes, which could encode novel proteins. Moreover, the dispensable chromosomes show extremely low synteny with other Dothideomycete genomes.

http://www.pri.wur.nl/UK/research/research+themes/Interaction+between+plants+pests+and+diseases/wheat/

Composition of Metagenomes from Yellowstone Hot Spring Microbial Mats Constructed by Photosynthetic Prokaryotes

Christian G. Klatt,[1] **Jason Wood**[1]* (montanawoody@gmail.com), Doug Rusch,[2] Donald A. Bryant,[3] William P. Inskeep,[1] and David M. Ward[1] (umbdw@montana.edu)

[1]Department of Land Resources and Environmental Sciences, Montana State University, Bozeman; [2]The J. Craig Venter Institute, Rockville, Maryland; and [3]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, State College

We have analyzed metagenomic data obtained from microbial mats constructed by cyanobacteria in alkaline siliceous hot springs and by anoxygenic phototrophic bacteria in sulfidic carbonate hot springs of Yellowstone National Park. Metagenomic libraries were analyzed from four samples—the top green layers of ~60-65°C cyanobacterial mats from Octopus and Mushroom Springs, the undermat layers from the 60°C Mushroom Spring mat, and mats from Bath Lake Vista Annex (BLVA) Spring taken at two different seasonal time points. We have used genomes for representative isolates cultivated from these or similar systems in BLAST recruitment analyses.

The majority of metagenomic sequences from the top green layers of cyanobacterial mats were recruited by genomes of phototrophic isolates (*Synechococcus* sp. A and B', *Roseiflexus* sp. RS1, *Chloroflexus aurantiacus* sp. J-10-fl, Y-400-fl and 396-1, *Candidatus* Chloracidobacterium thermophilum and *Chloroherpeton thalassium*), with much lower recruitment by genomes of nonphototrophic isolates. Examination of % nt identity of recruited sequences revealed that these mat layers contain populations that are: (i) very closely related (i.e., >90% nt identity) to *Synechococcus*, *Roseiflexus* and *C. aurantiacus* 396-1, (ii) more distantly related (70-90% nt identity) to *Candidatus* C. thermophilum, *C. thalassium*, and other *Chloroflexus* strains, (iii) so distantly related to reference genomes that recruitment may have been fortuitous and (iv) not related to any of the reference genomes. By aligning paired sequences recruited by the *Synechococcus* sp. A genome in the BLAST analysis against this genome as a function of % nt identity, we noted the presence of sequence subpopulations that are either extremely closely related or more distantly related to this reference genome. Synteny increased with relatedness of metagenomic sequences to genomic homologs. Sequences related to *recA* exhibited a similar view of metagenome composition with high representation of dominant phototrophs, but also indicated the presence of *Aquificales, Firmicutes,* and multiple divisions within the *Proteobacteria*.

An independent oligonucleotide composition analysis was done with scaffolds constructed with metagenomic sequences by the Celera assembler. Isolate genomes and scaffolds greater than 5kb in length were profiled using the frequency distribution of all possible tri-, tetra-, penta-, and hexa-nucleotides and these profiles were compared using a principal components analysis. Scaffolds with similar oligonucleotide-composition showed evidence of similar taxonomic origin, as scaffolds containing phylogenetic marker genes grouped with genomes of similar oligonucleotide character. Most scaffolds were grouped using a k-means clustering algorithm into clusters containing the isolate genomes used as references. Scaffolds grouping in clusters not containing reference genomes may represent uncharacterized community members that could be targeted for future analysis.

Additional metagenomic libraries from Mushroom Spring bottom layers were subjected to a similar BLAST recruitment analysis. These BLAST results indicated that the lower layers of this mat community are dominated by the presence of organisms closely related to *Roseiflexus* spp., with decreased representation of *Synechoccocus* spp. Metagenomic libraries originating from BLVA had very little representation from oxygenic phototrophs, and were dominated by sequences from *Chloroflexus* and *Roseiflexus* spp. Future work will focus on the differences in diversity patterns of anoxygenic phototrophs among these metagenomic libraries.

Soil Metagenomics and Carbon Cycling in Terrestrial Ecosystems: Climate Change Response at the DOE's FACE and OTC Research Sites

Cheryl Kuske* (kuske@lanl.gov), Gary Xie, John Dunbar, Larry Ticknor, Yvonne Rogers, Shannon Silva, Laverne Gallegos-Graves, Stephanie Eichorst, Shannon Johnson, Don Zak, Rytas Vilgalys, Chris Schadt, Dave Evans, Patrick Megonigal, Bruce Hungate, Rob Jackson, David Bruce, and Susannah Tringe

Los Alamos National Laboratory, Los Alamos, New Mexico

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. The DOE's long-term free air CO_2 enrichment (FACE) and open top chamber (OTC) field experiments will come to completion over the next two years, 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 sites are large, replicated field experiments that encompass forest, scrubland, desert, and wetlands, allowing comparison of belowground responses of very different terrestrial ecosystems to elevated CO_2.

Using a variety of comparative metagenomic sequencing approaches coupled with functional measurements, we are determining the effects of elevated CO_2 treatment on the soil microbial community at six of the DOE field sites. Taxonomic profiling of the bacteria, archaea, and fungi is being conducted to provide an overall assessment of microbial community structure and as a baseline for shotgun metagenomics comparisons. Suites of functional genes that are important in carbon and nitrogen cycling are being sequenced from replicate plots at each site to focus on populations involved in key processes. At two sites where we have evidence that soil fungi are involved in response to elevated CO_2 conditions, the soil fungal transcriptome is being sequenced. A broader, shotgun metagenomic approach has been initiated. We plan to also investigate the seasonal responses to elevated CO_2, and the interactive effects of ozone and addition of soil nitrogen using these approaches.

Genomics and Transcriptomics Studies on Mycelium of Shiitake Mushroom *Lentinula edodes* Growing in Lignocellulosic Media Using Sequencing-By-Synthesis Method

Iris S.W. Kwok* (kwokdorami@hotmail.com), Winnie W.Y. Chum, Tommy C.H. Au, and H.S. Kwan

Department of Biology, The Chinese University of Hong Kong, HKSAR, China

Lentinula edodes degrades lignocellulose efficiently. Various lignocellulolytic enzymes are produced during vegetative growth of mycelia. We are interested in identifying these lignocellulolytic enzyme-coding genes by means of (1) genome survey sequencing (GSS) and (2) large scale sequencing of cDNA of lignocellulose-grown mycelia.

Two GSS were performed, one was from *L. edodes* dikaryotic strain L54 and the other was from the monokaryotic parent of L54. Altogether, two sequencing runs provided about 153Mb genome sequence (approximately 3.5x coverage), in which 859,400 shortgun reads were obtained. From a single sequencing run of lignocellulose-grown mycelial cDNA, 42,377 reads were obtained. About 6000 contigs were assembled. All these cDNA contigs were annotated by BLASTX and categorized according to Gene Ontology. To obtain a more complete analysis on the variety of lignocellulolytic enzymes in *L. edodes*, all of the raw reads in GSS and cDNA sequencing, together with thousands of *L. edodes* ESTs available in NCBI, were subjected to BLAST against FOLy (Fungal Oxidative Lignin enzymes) and CAZy (Carbohydrate-Active enZymes) databases. Those having e-value $<10^{-5}$ were selected and classified into different FOLy and CAZy groups.

With the help of high-throughput sequencing, we may have better understanding on the lignocellulolytic system, especially the lignocellulolytic enzymes, of *L. edodes*.

Assembly Improvement on Data from New Sequencing Technologies

Kurt LaButti,[1] Brian Foster,[1] Stephan Trong,[2] Steve Lowry,[1] Cliff Han,[3] Tom Brettin,[3] Susan Lucas,[2] and **Alla Lapidus**[1]* (alapidus@lbl.gov)

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Lawrence Livermore National Laboratory, Livermore, California; and [3]Los Alamos National Laboratory, Los Alamos, New Mexico

The Joint Genome Institute (JGI) is a world wide leader in microbial genome sequencing. More than 300 bioenergy, bioremediation, and carbon sequestration related bacterial genomes have been completely sequenced at the JGI. Recently, we modified our microbial sequencing pipeline to accommodate the benefits of next generation sequencing platforms. Currently, we produce 15-20 x coverage from the titanium 454 sequencer, 10 x coverage of paired ends from titanium, and 50–100 x from Illumina GAii sequencer for each microbial project. Two new tools were developed to further automate the genome finishing process, one focusing on gap resolution and the other focusing on improving low quality consensus bases and correcting small insertions and deletions.

Data from the 454 platform and the Newbler assembler allow us to produce highly reliable skeleton of the microbial chromosome. Gaps in Newbler assemblies are usually caused by repeats (Newbler keeps those reads separately), very strong secondary structures and artifacts of the PCR process (specific for 454 paired end libraries). Some gaps in draft

assemblies can be closed by adding back reads from repeats. To facilitate gap closure, we created a tool that identifies read pairs belonging to a gap, assembles them using the PGA assembler in the form of a subassemblies, determines if gap is closed and then designs further lab experiments if more work is needed to complete the gap. This software package named Gap Resolution and was designed specifically to help automate the process of closing gaps in next generation assemblies.

Illumina data is used in our microbial finishing process to correct potential frame shifts and to improve low quality areas. Traditional polishing (quality improving step) is a time consuming and costly process. Significant improvents have been made to the process by using a tool that we call the Polisher. This tool was developed to automatically use Illumina data to improve quality and correct indels by aligning ultra short reads to the existing consensus. To get the best result from the Polisher, the tool removes low quality reads from the Illumina data set. In future work, we plan to further automate gap closure by using de novo short read assemblies.

This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

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] and Daniel van der Lelie[1] (vdlelied@bnl.gov)

[1]Brookhaven National Laboratory, Upton, New York; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]National Renewable Energy Laboratory, Golden, Colorado; and [4]Verenium Corporation, San Diego, California

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.

High-Throughput Genetic Dissection of Carbon Metabolism Networks in Two Divergent Aerobically Fermenting Yeasts

Tomas Linder* (tlinder@cmp.ucsf.edu) and Nevan Krogan

Department of Cellular and Molecular Pharmacology, University of California, San Francisco

The yeasts *Saccharomyces cerevisiae* (budding yeast) and *Schizosaccharomyces pombe* (fission yeast) share a long history in brewing tradition as well as a more recent history as popular genetic models for understanding the basic biology of eukaryotic cells. Although both species are ascomycete fungi, estimates of the time of their evolutionary divergence ranges from 500 million years ago to around 1 billion years ago. Yet these two divergent species also share the unusual ability to ferment sugars in the presence of oxygen (the so-called Crabtree effect), which makes them attractive from a biotechnological perspective. Phylogenetic data strongly suggests that these two species developed this ability independently of each other and it generally thought that this ability co-evolved with the first appearance of sugar-rich fruit in angiosperm plants during the Cretaceous. A comparative study of the carbon metabolism networks of these two Crabtree positive yeasts will help determine the genetic basis for aerobic fermentation in fungi.

To date we have a fairly comprehensive overview of the basic yeast carbon metabolism network and how it regulates the conversion of carbon substrates into energy and biosynthetic precursors. However, it remains to be determined (1) what the complete sets of genes are that are required for growth on any one particular carbon substrate and (2) how and where each of these genes fit into the larger metabolic network. We have addressed the first question through agar array-based phenomic screens of haploid cells from single gene deletion libraries in both *S. cerevisiae* and *S. pombe* to identify all genes with carbon substrate-specific phenotypes on a number of different fermentative and respiratory carbon substrates.

We are currently placing these genes in the context of the overall carbon metabolism network in both *S. cerevisiae* and *S. pombe* through systematic high-throughput generation

of double mutants in haploid yeast cells. To this end we have employed the synthetic genetic array (SGA) technology, which allows for the rapid and systematic analysis of synthetic lethal and suppressive genetic interactions by the simultaneous generation of tens of thousands of double deletion mutants. A haploid strain lacking a specific gene (the query strain) is crossed to a library of haploid single deletion strains of all non-essential genes arrayed on agar plates. After mating and sporulation, engineered genetic markers allow for the selection of haploid cells carrying both the query deletion and the corresponding library deletion. Subsequent comparison of carbon substrate-specific phenotypes between single and double deletion mutants allows for the identification of gene redundancy as well as the detection of novel bypass pathways in both the carbon metabolism network and its associated regulatory network.

This technology has the potential to greatly streamline the generation of yeast strains for industrial applications. The SGA methodology can rapidly identify deletions that would improve industrial performance of a particular strain, whether it is fermentation efficiency, ethanol tolerance, the ability to ferment novel substrates or the production of fermentation products other than ethanol.

High Throughput Genomic Sequencing Reveals Pericentromeric Clustering of Sperm MARs

Amelia K. Linnemann[1]* (amelia@compbio.med.wayne.edu), Adrian E. Platts,[1,2] Claudia Lalancette,[2] Norman Doggett,[3] Douglass T. Carrell,[4] and Stephen A. Krawetz[1,2,5]

[1]The Center for Molecular Medicine and Genetics, and [2]Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, Michigan; [3]Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico; [4]Andrology and IVF Laboratories, Department of Surgery (Urology) and Physiology, University of Utah, Salt Lake City; and [5]Institute for Scientific Computing, Wayne State University School of Medicine, Detroit, Michigan

During spermiogenesis, the majority of histones are replaced by the highly basic arginine-rich, cysteine-containing protamine proteins. This can exceed 98% in mouse and 85% in man, resulting in a compaction of the genome to 1/14th of the size observed in somatic cells. The highly compacted mature sperm genome is transcriptionally quiescent. Somatic genes generally appear silenced when intragenic regions are attached to the nuclear matrix at MARs, or matrix attachment regions. Accordingly, it has been proposed that within this context, nuclear matrix attachment in sperm may serve a similar genomic silencing function. To test the tenet, MARs on human chromosomes 14 – 18 were identified using genomic array CGH (comparative genomic hybridization). Comparison to multiple somatic cell types revealed a subset of regions constitutively attached to the nuclear matrix. These can be differentiated as sites of attachment that are specific to somatic cells and those that appear only in sperm cells. Analysis of sperm MARs by 454 and Illumina GA2 sequencing revealed sites of attachment along the length of each chromosome. The sites of marked enrichment appear at the pericentromeric regions of a subset of chromosomes within the human genome. These are asymmetrically distributed such that only the short arm or the long arm of any given chromosome displays this extended attachment. This suggests that nuclear matrix mediated chromosomal organization in sperm may provide boundaries to structurally segregate chromosome arms from the centromeres. Further analysis is underway to determine if this segregation protects specific groups of genes near the pericentromeric enriched regions that may play a role in either transcription throughout spermatogenesis or after fertilization.

Acknowledgements: This work was supported in part by a grant to ND and SAK from the JGI and in part by National Institutes of Health grant HD36512 and the Wayne State University Research Enhancement Program in Computational Biology to SAK.

Marine Bacterial Alkaline Phosphatase Genes a Proxy for P-Stress

Richard A. Long* (rlong@biol.sc.edu) and Haiwei Luo

Department of Biological Sciences and Marine Science Program, University of South Carolina, Columbia

Alkaline phosphatase activity has commonly been used as an indicator for phosphorus stress in the ocean. Marine bacteria excrete alkaline phosphatases that hydrolyze phosphoesters, the most abundant phosphorous-compounds in the ocean. We systematically examined the distribution of alkaline phosphatase genes (phoA, phoD and phoX) in two oligotrophic microbial metagenomes from the Sargasso Sea and the North Pacific Subtrophical Gyre (NPSG). We found that the normalized alkaline phosphatase gene occurrence is significantly higher in the surface waters of the Sargasso Sea than that of the NPSG. This concurs with previous chemical studies that suggest while both systems are phosphorus stressed, the concentration of phosphate in the Sargasso Sea is two orders of magnitude lower than the NPSG. Our analysis of alkaline phosphatase genes suggests that *Alteromonadales* potentially has an important role in dissolved organic phosphorous remineralization in oligotrophic oceans. In addition, *Bacteroidetes*, *Silicibacter*, *Prochlorococcus*, and *Synechococcus* may also be important. Most eco-physiological studies have focused upon the monoesterase activity of alkaline phosphatase (phoA); however, we saw a high occurrence of diesterase cleaving alkaline phosphatases (phoD and phoX) , which suggests that hydrolysis of phosphodiesters may play a significant, but unrecognized, role in phosphorus cycling in the ocean.

Characterizing the Transcriptional Space of Loblolly Pine (*Pinus taeda* L.)

W. Walter Lorenz[1]* (wlorenz@uga.edu) and Jeffery F.D. Dean[1,2]

[1]Warnell School of Forestry and Natural Resources and [2]Department of Biochemistry and Molecular Biology, University of Georgia, Athens

Loblolly pine is the single most important forest tree species grown commercially in the United States. It is native to the southeastern U.S. where it is used as the primary feedstock for pulp and paper, as well as the dimensional lumber processing industries. Loblolly pine also has potential as a feedstock source for biofuels production. Given its commercial importance and economic potential, better understanding of the genetics that underpin the growth characteristics, wood quality, and response to biotic and abiotic stresses in loblolly pine will be essential for the continued improvement of this resource. Like the rest of the conifers, the *P. taeda* genome, at ca. 2.0 x 10^{10} bp, is large – roughly seven times the size of the human genome. Thus, sequencing of the complete genome is unlikely in the near future. Consequently, genome characterization efforts to date have focused on gene discovery at the transcriptional level. A Community Sequencing Project has been initiated

at the JGI to increase the depth of coverage of the loblolly pine transcriptome and also provide transcriptome sequence information for additional conifer species representing clades that have not previously received study at the genomic level. In a pilot study to examine the effects of normalization on gene family representation, normalized (CFCN) and non-normalized (CFCP) cDNA libraries were synthesized from RNA isolated from first flush candle tissues (elongating apical shoot tips) pooled from three loblolly genotypes: CCLONES 40430, 40368, and 41586. *De novo* transcriptome assemblies of both the individual and combined library datasets were performed using the SeqMan NGen assembler and the results were compared to assemblies generated by the JGI bioinformatics pipeline as well as the miraEST assembler. Additional assemblies were generated by combining the GS-FLX datasets with the Sanger-based ESTs available in GenBank, and also by assembling the new sequence data to templates obtained from the unigene set identified at NCBI for loblolly pine. Future plans for the CSP project include production of an additional 1.5 Gb of loblolly pine cDNA sequence information using the Roche GS-FLX Titanium platform, which should begin to give us fairly comprehensive picture of the depth and complexity of the transcriptome in this important gymnosperm species.

CoGe: Making Comparative Genomics Easy

Eric Lyons* (elyons@nature.berkeley.edu)

University of California, Berkeley

http://synteny.cnr.berkeley.edu/CoGe

The biological research community has entered the genomics age. Although there are many freely available online software solutions for obtaining, comparing, and visualizing genomic data, these resources suffer four major limitations. First is the disparate nature of obtaining data where researchers often have to navigate to multiple websites because each genomics platform provides accesses to a limited subset of all publicly available genomic information, and in many cases, only to a single genome, all of which change as genomic data is periodically updated. Second, sequence analysis tools differ widely in their ability to detect particular patterns of sequence similarity, are likewise dispersed across multiple websites, and in some cases are only available for online use against a limited number of genomes. This requires researchers to not only navigate to multiple locales to analyze their sequence data, but sometimes reformat their data in order for it to be properly used by an algorithm. Third, and often neglected, is the ability to easily visualize sequence data and their comparisons in order to glean patterns of change. Fourth is the ability to iteratively refine an analysis in order to expand the amount of genomic sequence analyzed, hone in on a particular region of interest, or change sequence comparison algorithms in order to detect different patterns of sequence similarity. Importantly, interlinking these four steps has not been accomplished to make this process efficient.

CoGe represents a web-based software system that addresses all of these limitations facing comparative genomics. Using CoGe, many patterns in the evolution of genomes may be characterized including synteny, whole genome duplication events, fractionation, gene deletion events, local gene duplications, inversions, translocations, mis-annotations, and conserved noncoding sequence. CoGe currently stores genomes from over 5,500 organisms. CoGe's ability to allow researchers anywhere in the world to rapidly identify genes and genomic regions of interest and visualize their evolution forges a powerful new tool for any biologist.

CoGe is publicly available at: http://synteny.cnr.berkeley.edu/CoGe

Genome Sequencing of 1,4-Dioxane-Utilizing Bacterium *Pseudonocardia dioxanivorans* CB1190

Shaily Mahendra,[1] **Christopher M. Sales**[2]* (chris.sales@berkeley.edu), Rebecca E. Parales,[3] and Lisa Alvarez-Cohen[2]

[1]Rice University, Houston, Texas; [2]University of California, Berkeley; and [3]University of California, Davis

Pseudonocardia dioxanivorans CB1190 was characterized with respect to its morphology, physiology, and phylogeny. CB1190 is an unusually important bioremediation strain because it can grow using 1,4-dioxane as its sole source of carbon and energy. Strain CB1190 exhibits the fastest reported rates of 1,4-dioxane degradation and the capability to degrade other toxic pollutants such as benzene, toluene, tetrahydrofuran, and methyl butyl ether. CB1190 can also fix dinitrogen, which makes it an attractive bioaugmentation culture for nitrogen-limited environments.

We are currently collaborating with the Department of Energy's Joint Genome Institute (JGI) to sequence the whole genome of strain CB1190 in order to facilitate future studies of this resilient biodegrading organism. Results to date confirm the identity of the strain and its phylogenetic position. The genome size is 6.4 Mb and G+C content is 72 mol%. A shotgun sequence with 8x coverage of the complete genome is currently under production at JGI. Automated assembly and draft annotation of the sequence will be performed in collaboration with JGI. Manual curation of the genome will involve tasks such as identifying genes missed by JGI's automated gene prediction pipeline, as well as finding overlapped genes and fixing lengths of genes whose sizes were incorrectly predicted. The genome sequence will then be used to generate a whole-genome microarray to study expression of genes under a variety of growth conditions and environmental disturbances. Significantly differentially expressed genes found by the microarray will be validated using reverse transcriptase quantitative PCR (RT-qPCR).

Identification of Genes Involved in Acetylation of Cell Wall Polysaccharides in *Arabidopsis thaliana*

Yuzuki Manabe* (ymanabe@lbl.gov), Andreia Michelle Smith, Caroline Orfila, Chithra Manisseri, Özgül Persil Çetinkol, Brad Holmes, Joshua Heazlewood, and Henrik Vibe Scheller

Joint BioEnergy Institute

Acetylation of cell wall polysaccharides has long been observed in various plant species. However, the enzymes involved in the acetylation have thus far not been identified in plants. While the *in vivo* role of polysaccharide acetylation is still unclear, it is known to affect biofuel yield from lignocellulosic biomass due to inhibition of the microorganisms. Therefore decreasing acetylation levels in lignocellulosic biomass may increase the efficiency of biofuel production. We have analyzed four *Arabidopsis* homologues of a protein known to be involved in polysaccharide acetylation in a fungus. *Arabidopsis* mutants with insertional mutagenesis in the respective genes were identified, and we found that at least one of the mutants, designated *reduced wall acetylation (rwa1, rwa2, rwa3*

and rwa4) had decreased levels of acetylated cell wall polymers. Two independent alleles of *rwa2* mutants were examined by analyzing alcohol insoluble residues extracted from leaves. Extracts treated with 0.1M NaOH released about 20% lower amounts of acetic acid when compared to wildtype. Interestingly, the composition of the cell wall monosaccharides in *rwa2* was not altered. There was no apparent visible difference observed between wildtype and either allele of mutants at any developmental stages. Results of experiments to determine the specific polymers affected in the *rwa2* mutants will be reported.

Hydrogenobaculum Population Genomics: Linking Phylogeny, Geochemistry, Genetics, and Ecologic Function

Timothy R. McDermott[1]* (timmcder@montana.edu), Scott Clingenpeel,[1] and Scott Miller[2]

[1]Thermal Biology Institute, Montana State University, Bozeman; and [2]Division of Biological Sciences, University of Montana, Missoula

Background: *Hydrogenobaculum* inhabiting the Yellowstone geothermal complex are the focus of a population genomics study. Several pure culture isolates that are phylogenetically 100% identical (full-length 16S rDNA sequence and ITS sequence), but that differ in their ability to grow with H_2 and or H_2S as an energy source are being genome sequenced. This effort is in parallel with a metagenomic study of a mat community in Dragon Spring (our primary study site and a NSF Microbial Observatory) where *Hydrogenobaculum* is a dominant member (99% of *Bacteria* PCR clones). Based on extensive chemical analysis of the spring, we know that chemolithoautotrophy is a highly relevant metabolism in this spring, which is chemostat-like in nature with respect to temperature, pH, and flux of H_2 and H_2S. Importantly, however, this habitat is also comprised of overlapping temperature and geochemical gradients that provide a continuum of niche opportunities that theoretically could support the maintenance of various genetic alterations that might account for the variation we have observed in ability to utilize H_2 or H_2S.

Results: Draft genome coverage of one isolate is being examined for specific functions of interest and metagenome reads are also being studied. Genome sequence revealed novel sequences coding for arsenite oxidases and that have been used for expression and diversity analysis. These and other current developments will be summarized. This work has been supported by the NSF Microbial Observatories and DOE-JGI-Community Sequencing programs, and the NASA-supported Thermal Biology Institute.

Pipeline for Novel Biomass Degrading Enzymes

David Mead* (dmead@lucigen.com), Becky Hochstein, Julie Boyum, Cate Brumm, Mai Lee, Sarah Vande Zande, Eric Steinmetz, Ronald Godiska, Krishne Gowda, and Phil Brumm

Lucigen Corp., Middleton, Wisconsin; and C56 Technologies Inc., Middleton, Wisconsin

Having the correct enzymes for biomass degradation, at a low enough price to be affordable, is a major goal of biofuels research. Currently, the biomass-degrading enzyme products that are commercially available are too expensive for practical use in the

production of biofuels. The discovery of new high specific activity biomass active enzymes for evaluation in degradation studies is the focus of this research. An improved pipeline for enzyme discovery specific to the problems unique to this field was developed and validated, and a number of new carbohydrases were produced. Endo- and exo-cellulases and hemicellulases were discovered with high specific activity and broad specificity. We have over expressed, purified and characterized a number of unique cellulytic enzymes and will present data on representative examples. The next step is to develop a minimal set of biomass active enzymes that eliminates the bottleneck in cellulose degradation, in conjunction with research scientists at the Great Lakes Bioenergy Research Center at the University of Wisconsin, Madison.

Whole-Transcriptome Sequencing of an Interspecific Hybrid of *Eucalyptus* Using Illumina mRNA-Seq: Preliminary Assembly and Challenges

Eshchar Mizrachi,[1] Charles Hefer,[2] Martin Ranik,[1] Jean-Marc Celton,[3] D. Jasper G. Rees,[3] Fourie Joubert,[2] and **Alexander A. Myburg**[1]* (zander.myburg@fabi.up.ac.za)

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI) and [2]Unit for Bioinformatics and Computational Biology, Department of Biochemistry, University of Pretoria, South Africa; and [3]Department of Biotechnology, University of the Western Cape, Cape Town, South Africa

Fast-growing hybrids of *Eucalyptus* tree species are widely used in clonal plantation forestry in countries such as South Africa and Brazil. Eucalypt hybrids are highly productive fibre crops and some of the most prolific producers of lignocellulosic biomass known. Despite ongoing efforts to generate genomic resources for *Eucalyptus* (including the JGI *Eucalyptus grandis* Genome Project and several genetic mapping and EST sequencing projects), little is known about the dynamics (sequence, structure and expression) of the transcriptomes of eucalypt hybrids. Of interest are the expression levels and structural (splice) variation of mRNA transcripts in wood-forming (and other) tissues of hybrid trees. Allele-specific expression patterns may also underlie the unique properties of interspecific hybrids. Deep mRNA sequencing using ultra-high-throughput Illumina mRNA-Seq technology is a potential approach to answering these questions. We present preliminary assembly data for the transcriptome of an F1 *E. grandis* x *E. urophylla* hybrid clone grown in South African plantations. The total raw Illumina data thus far consists of 124 million paired-end (PE) reads (approximately 6.9 Gb). Auxiliary data available to guide the assembly include publicly available Sanger and 454 EST sequences, as well as a 4.5X draft assembly of the *E. grandis* genome sequence (JGI). Some of the challenges discussed include variation of length of paired-end data (36, 55 and 60 base paired-end reads), assembly algorithms and varying levels of ribosomal RNA in the library preparation. In addition, estimation of allowable error rate for assembly is complicated by the fact that the transcriptome is that of a hybrid (effectively that of two different species) and that the reference genome is of only one parental species. However, the combination of Illumina PE and mRNA-Seq technology has allowed us to obtain an estimated 30X coverage (1.8 Gbp of mappable sequence) of the *Eucalyptus* transcriptome (assuming a maximum of 30,000 expressed genes of on average 2000 bp). Further increases in sequence coverage should allow us to assemble the near-complete transcriptome of the *Eucalyptus* hybrid clone and to identify SNP markers for genetic mapping of candidate genes in F2 hybrid progeny.

Evolution of Nodulating Opportunists Threatens Legume Productivity

Kemanthi G. Nandasena* (kemanthi@murdoch.edu.au), Ravi P. Tiwari, Graham W. O'Hara, Wayne G. Reeve, and John G. Howieson

Centre for *Rhizobium* Studies, Murdoch University, Western Australia

Legumes and their root nodule bacteria, commonly known as rhizobia, are often introduced to managed agricultural ecosystems to improve organic fertility, their nitrogen economy and farming systems flexibility. Rhizobia play a significant role in agricultural systems owing to their ability to transform atmospheric nitrogen into forms usable by leguminous plants. *Biserrula pelecinus* L. is one of only three deep rooted annual legume species which is widely used in commerce with the potential to reduce the development of dryland salinity in Australia. Furthermore, this species is growing well in acidic, duplex soils of Australia due to the interaction with an acid tolerant rhizobia and is also rapidly gaining popularity across Australia due to many of its other beneficial agronomic attributes, which include drought tolerance, hard seed production, grazing tolerance, high seed yield, easy harvesting characteristics and insect tolerance. *B. pelecinus* L. is native to the Mediterranean basin and is nodulated by *Mesorhizobium ciceri* bv. *biserrulae*, a rhizobial species naturally absent from Australia. This symbiotic interaction is highly specific and previous studies have shown that Australian rhizobial populations are not capable of nodulating *B. pelecinus*.

Successful production of a legume depends on the efficiency of N_2 fixation by its rhizobia. Competition for nodulation occurs whenever more than a single rhizobial genotype capable of nodulating the legume is resident within the soil. The multi billion dollar asset attributed to symbiotic nitrogen fixation in Australia is often threatened by nodulation of legumes by rhizobia that are ineffective or poorly effective in N_2 fixation.

This study investigated the development of rhizobial diversity for *B. pelecinus*, six years after its introduction, and inoculation with *M. ciceri* bv. *biserrulae* strain WSM1271, to Western Australia. Molecular fingerprinting of 88 nodule isolates indicated that seven were distinctive. Two of these were ineffective while five were poorly effective in N_2 fixation on *B. pelecinus*. Three novel isolates had wider host ranges for nodulation than WSM1271. These ineffective or poorly effective strains had distinct carbon utilization patterns and the sequencing of the housekeeping genes *16S rRNA*, *dnaK* and *GSII* revealed that they belong to two new species within the bacterial genus *Mesorhizobium*. The ineffective strains have been named *M. opportunistum*, while the poorly effective strains *M. australicum*. Localization and sequencing of the symbiotic genes in these novel species as well as in WSM1271 revealed the symbiotic genes are located on a symbiosis that has been transferred from WSM1271 to the novel isolates enabling them to nodulate *B. pelecinus*. This is the first evidence for the evolution of opportunistic rhizobia with inferior N_2-fixing ability following *in situ* transfer of symbiotic genes on a symbiosis island from an inoculant strain to other soil bacteria.

Deep Transcriptional Profiling to Understand Triacylglycerol Production in Oil Producing Tissues

John Ohlrogge[1] (Ohlrogge@msu.edu), Tim Durrett,[1] Adrian Troncosco,[1] Jilian Fan,[1] Xia Cao,[1] Erika Lindquist,[2] Christa Pennacchio,[2] and Curtis Wilkerson[1]

[1]Department of Plant Biology and DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing; and [2]DOE Joint Genome Institute, Walnut Creek, California

Unlike microarrays, EST sequencing provides a method of transcript analysis that allows quantitative comparisons between genes and between different plant species. In order to identify genes involved in plant oil biosynthesis and the transcription factors and other regulatory systems that control oil accumulation, Michigan State, together with JGI has sequenced over 10 million ESTs from a variety of oilseeds and other oil rich tissues. Why do we need millions of ESTs? Key enzymes of lipid metabolism (e.g. acyltrasferases, phospholipases, thioesterases) are very low abundance and can be difficult to detect by conventional EST sequencing. Deep EST sequencing using 454 pyrosequencing provides a large increase in EST sequence information and allows us to accurately quantify low level expression. By sequencing libraries from multiple species we obtain information on what similarities and differences distinguish oil synthesis in seeds producing unusual fatty acids and in seeds compared to other tissues such as mesocarp that produce high oil levels.

Replicate analysis of samples (including cDNA synthesis and PCR) gave a 0.99 correlation coefficient between # reads per gene. Therefore, 454 sequencing is technically and biologically reproducible and provides an accurate measure of gene expression. We have observed that core enzymes of fatty acid biosynthesis are, in general, expressed in consistent stoichiometric ratios in a number of different oilseeds and tissues. Therefore, those genes that fall outside the usual stoichiometry offer insight into unique metabolism. For example, we observe very low expression of the FatB thioesterase that controls saturated fatty acid production in castor, which agrees with the fact that castor is an oilseed with extremely low saturated fatty acid content. In addition: 1) 18:0-ACP desaturases are most abundant transcript of plastid FA metabolism. 2) DAGAT2 transcripts are 100 fold > than DAGAT1 in castor. 3) LPCAT transcripts are next most abundant acyltransferase followed by LPAT and “GPAT9”. High LPCAT is consistent with our recent flux model for lipid synthesis in soybean.

Discovery of a gene encoding an enzyme catalyzing formation of acetyl-glycerides. Endosperm tissue from *Euonymus alatus* (Burning Bush) accumulates unusual triacylglycerols (TAGs) with an *sn*-3 acetate group instead of a long-chain fatty acid. Because of their low viscosity these unusual acetyl-glycerides (ac-TAGs) TAGs have added value applications in direct use as biodiesel and lubricant oil feedstocks. In addition to producing ac-TAGs, Euonymus fruit also synthesize normal, long-chain TAGs (lc-TAGs) in their aril tissue. The close developmental coordination and spatial proximity of two tissues with the ability to produce different TAGs presents a unique opportunity to understand the accumulation of unusual TAGs in plants. By sequencing ESTs from these tissues we have identified candidate genes involved in ac-TAG biosynthesis. One candidate expressed in endosperm but not aril tissue produces ac-TAG when expressed in yeast. In addition, microsomes of the yeast catalyze ac-TAG biosynthesis from acetyl-CoA and DAG.

Scaling Up the 454 Titanium Library Construction and Pooling of Barcoded Libraries

Wilson Phung* (wphung@lbl.gov), Chris Hack, Harris Shapiro, Susan Lucas, and Jan-Fang Cheng

DOE Joint Genome Institute, Walnut Creek, California

We have been developing a high throughput 454 library construction process at the Joint Genome Institute to meet the needs of de novo sequencing a large number of microbial and eukaryote genomes, EST, and metagenome projects. We have been focusing efforts in three areas: (1) modifying the current process to allow the construction of 454 standard libraries on a 96-well format; (2) developing a robotic platform to perform the 454 library construction; and (3) designing molecular barcodes to allow pooling and sorting of many different samples. In the development of a high throughput process to scale up the number of libraries by adapting the process to a 96-well plate format, the key process change involves the replacement of gel electrophoresis for size selection with Solid Phase Reversible Immobilization (SPRI) beads. Although the standard deviation of the insert sizes increases, the overall quality sequence and distribution of the reads in the genome has not changed. The manual process of constructing 454 shotgun libraries on 96-well plates is a time-consuming, labor-intensive, and ergonomically hazardous process; we have been experimenting to program a BioMek robot to perform the library construction. This will not only enable library construction to be completed in a single day, but will also minimize any ergonomic risk. In addition, we have implemented a set of molecular barcodes (AKA Multiple Identifiers or MID) and a pooling process that allows us to sequence many targets simultaneously. Here we will present the testing of pooling a set of selected fosmids derived from the endomycorrhizal fungus *Glomus intraradices*. By combining the robotic library construction process and the use of molecular barcodes, it is now possible to sequence hundreds of fosmids that represent a minimal tiling path of this genome. Here we present the progress and the challenges of developing these scale-up processes.

Brachypodium distachyon Transcriptomics

Henry D. Priest[1]* (priesth@onid.orst.edu), Samuel Fox,[1] Scott A. Givan,[1] Sergei Filichkin,[1] Todd P. Michael,[2] and Todd C. Mockler[1]

[1]Department of Botany and Plant Pathology and Center for Genome Research and Biocomputing, Oregon State University, Corvallis; and [2]Waksman Institute of Microbiology, Rutgers University, Piscataway, New Jersey

We are using hypothesis-generating approaches to lay the foundations for elucidating details of the transcriptional network in the model grass Brachypodium distachyon, a model for temperate grasses and bioenergy crops. We used Illumina (Solexa) sequencing to sample a collection of cDNA libraries representing a diverse array of tissues, treatments and developmental stages. The resulting EST data were aligned to the Brachypodium genome and used to assemble transcriptional units, including alternative splice variants. The high depth of sequencing and broad unbiased coverage available from the Illumina platform increases the chance of identifying low abundance transcripts. Our analysis provides a comprehensive view of the *Brachypodium* transcriptome and facilitates annotation efforts. We are using our empirical annotation of the transcriptome to aid design of a versatile oligonucleotide microarray platform that includes exon scanning and genome tiling features. We will use these arrays to generate a *Brachypodium* expression

atlas comparing tissues over development, diurnal and circadian time-courses, and stress conditions. This atlas will provide a hypothesis-generating foundation for elucidating the transcriptional networks underlying traits of major importance economically important crops including potential bioenergy grass crops.

Expansion of Signal Pathways in the Ectomycorrhizal Fungus *Laccaria bicolor*—Evolution of Protein Kinases and RAS Small GTPases

Balaji Rajashekar, Annegret Kohler, Tomas Johansson, Francis Martin, Anders Tunlid, and **Dag Ahrén*** (dag.ahren@mbioekol.lu.se)

Lund University, Sweden

The ectomycorrhizal fungus *Laccaria bicolor* has the largest genome of all fungi yet sequenced. The large genome size is partly due to an expansion of gene family sizes. Among the largest gene families are protein kinases and RAS small GTPases, which are key components of signal transduction pathways. Comparative genomics and phylogenetic analyses were used to examine the evolution of the two largest families of protein kinases and RAS small GTPases in *L. bicolor*. Expression levels in various tissues and growth conditions were inferred from microarray data. The two families had a large number of young duplicates (paralogs) that have arisen in the *Laccaria* lineage following the separation from the saprophyte *Coprinopsis cinerea*. The protein kinase paralogs were dispersed in many, small clades and a majority of them were pseudogenes. In contrast, the RAS paralogs were found in three large groups of RAS1-, RAS2- and RHO1-like GTPases with few pseudogenes. Duplicates of protein kinases and RAS small GTPase have either retained, gained or lost motifs found in the coding regions of their ancestors. Frequent outcomes during evolution are the formation of pseudogenes (nonfunctionalization) or proteins with novel structures and expression patterns (neofunctionalization).

Key words: ectomycorrhiza, gene duplications, gene family evolution, *Laccaria bicolor*, protein kinases, RAS small GTPases.

The Development of *Chlamydomonas reinhardtii* Chloroplasts as a Biotechnology Platform for the Production of Recombinant Proteins

Beth A. Rasala* (brasala@scrippps.edu), Andrea L. Manuell, Miller Tran, Machiko Muto, and Steven P. Mayfield

Department of Cell Biology, The Scripps Research Institute, La Jolla, California

Chlamydomonas reinhardtii is a eukaryotic green algae that has served as a genetic workhorse and model algae for understanding everything from the mechanisms of light- and nutrient-regulated gene expression to the assembly and function of flagella. Recently, we have genetically engineered the chloroplast of *C. reinhardtii* to express over 20 recombinant proteins, most of which were soluble, correctly folded, and bioactive. Thus, eukaryotic algae offer the potential to become an important and versatile platform for the economic production of recombinant proteins for medical, industrial, and scientific applications (Mayfield, et al., 2007). Additionally, the ability to genetically engineer algae could prove advantageous for biofuel accumulation and production. However, to date the

technologies for the expression of heterologous genes and the purification of recombinant proteins from algae lag behind other expression platforms. There are three main factors, acting alone or in concert, that limit recombinant protein accumulation in the micro-algae *Chlamydomonas reinhardtii*. First, poor accumulation of some recombinant mRNAs has been observed (Barnes, et al., 2005) that may result in lower levels of recombinant protein synthesis. Second, it is clear that heterologous mRNAs are translated much less efficiently then endogenous mRNAs, resulting in reduced protein accumulation. Finally, we have observed that some recombinant proteins are quite susceptible to degradation by chloroplast-localized proteases, and this is again limiting recombinant protein accumulation. Our goal is to develop micro-algae as an efficient biotechnology platform for the production of recombinant proteins by elucidating the molecular mechanisms of gene expression in alga chloroplast.

A Temporal, Paired Host and Viral Metagenomic Study to Investigate Host and Viral Population Dynamics and Diversity Within Acidic Yellowstone Hot Spring Environments

F.F. Roberto* (francisco.roberto@inl.gov), M.M. Bateson, A.C. Ortmann, T. Douglas, and M.J. Young

Idaho National Laboratory, Idaho Falls

The relatively low biological complexity of high temperature (>80°C) low pH (<4.0) environments in Yellowstone National Park hot springs provides an ideal opportunity to examine total viral and host diversity using metagenomic approaches. Our efforts have focused on two hot springs (CHANN041, NL10), each of which are dominated by relatively few novel archaeal members and their associated viruses. For each of these sites we are performing a temporal, paired host and viral metagenomic study. Viral and host samples have been collected over a 12 month period and are being sequenced by JGI. Initial results suggest that (i) each host spring is dominated by as few as a single (NL10, metagenome YNP2) or two (CHANN041, metagenome YNP1) archaeal species, (ii) it is likely that a complete or near complete assembly of a consensus genome of the dominant novel archaeal species present in NL10 will be possible directly from the metagenome using a combination of Sanger and 454 pyrosequencing technology, (iii) complete or near complete genomes of novel viruses are being assembled from the viral metagenomic data, and (iv) viral populations are highly dynamic over time in each of the hot springs.

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, Washington

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. 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. 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.

Plasticity and Genotype by Environment Interaction as Tools to Select for Growth Rate Trait Different Strains of *Pleurotus ostreatus* var. *Florida* (N001)

F. Santoyo Santos* (patximotxo@yahoo.es), A.G. Pisabarro, and L. Ramírez

Public University of Navarre, Department of Agrarian Production, Pamplona, Spain

Plasticity (P) and Genotype by Environment Interactions (GEI) reveal the performance of genotypes in different environments. We have analysed the influence of different temperatures on the growth rate of dikaryons growing on Petri dishes. Dikaryons were arranged in families formed obtained after the mating of compatible full-sib monokaryons meiotic-derived from *Pleurotus ostreatus*. Three different dikaryon families were constructed: two formed by dikaryons carrying un-recombined chromosome VIII versions (GrGr, fast growing; grgr, slow growing), and the other formed by dikaryons heterozygous for chromosome VIII (heterozygous Grgr). T-test analysis showed that no significant differences exist in the mean growth rates between GrGr and Grgr families at the

temperatures tested. Significant differences exist, however, between this group and that formed by dikaryons with the grgr genotype. Analysis of the results showed that the gene action governing the growth rate trait in full-sib dikaryons is dominant with slight differences at different temperatures. The effect of the environment on the growth rate trait (P) at the whole population as well as the identification of the genotypes with good performance in different environments (GEI) showed us that: high significant phenotypic plasticity (LOD score = 4,1) mapped to chromosome IV, just on in the same position where the second most important QTL controlling growth rate is located. This could be due to differences in allelic sensitivity of these loci to the temperatures. GEI was detected (with high significance) in several regions of the genome where QTLs were not previously identified. This could refute the gene regulation model that posits that specific loci may enhance (or suppress) expression of other genes in an environment-specific fashion. These results will help us to select those genotypes with best performance in each environment to be used in future experiments dealing with pre-treatment culture conditions of straw aimed at bio-ethanol production.

Exploring Variation Detection within a Wide-Range of Bioenergy-Relevant Species via Short Read Technology

Wendy Schackwitz* (wsschackwitz@lbl.gov), Joel Martin, Sirisha Sunkara, Mary Ann Pedraza, Maria Shin, David Hillman, Anna Lipzen, Crystal Wright, Feng Chen, and Len A. Pennacchio

DOE Joint Genome Institute, Walnut Creek, California

One major application of ultra-short read sequencing platforms is variation detection within species. Here we describe our efforts to identify single nucleotide substitutions and short indels in a diverse group of species relevant to bioenergy production. To date, we have completed resequencing studies using the Illumina GAii platform on 24 individual genomes covering inbred laboratory strains as well as wild isolates, ranging in complexity from haploid prokaryotes and fungi to diploid plants. As an initial foray into variation detection, we examined five finished microbes to assess the quality of the reference as well as our false-positive and false-negative discovery rates. Remarkably, no differences were identified between the reference genome sequence and Illumina data at >50X coverage suggesting perfect single base pair accuracy of the Sanger completed genome as well as a false-positive rate of zero. Using simulated variants, we estimate our false-negative rate at less than 3%. Next, we examined two strains of the more complex diploid genome *Arabidopsis thaliana* at ~20X coverage. Over 300k potential variants were identified in each strain with an estimated false-negative rate of ~20%. Finally, we have begun a project to sequence 10 ecotypes of the 500Mb outbred *Populus trichocarpa* genome to provide tools for quantitative trait loci mapping. In the initial ecotype at a depth of 25X we have identified over 800k potential variants, providing a marker on average every 500bp with an estimated false-negative rate of ~30%. These data offer additional evidence that ultra short read sequencing is a cost-effective way to identify variation for downstream genetic studies that define microbial directed evolution, catalogue feedstock diversity, and enable numerous bioenergy-driven biological investigations.

The data generated in this research will address the goals of developing bioremediation strategies and monitoring applications by identifying functional processes and ecological interactions at the biomolecular level that are crucial for effective bioremediation process performance. Members of the genus *Pseudonocardia* have been shown to degrade a wide variety of environmental contaminants, including those found at DoD and DOE sites.

P. dioxanivorans CB1190 can rapidly degrade 1,4-dioxane, which is a common co-contaminant with chlorinated solvents at DOE sites.

Eukaryotic Genome Projects at JGI-HudsonAlpha

J. Schmutz* (jschmutz@hudsonalpha.org), J. Grimwood, JGI-HA Group Members, and R.M. Myers

Joint Genome Institute–HudsonAlpha Genome Sequencing Center, HudsonAlpha Institute of Biotechnology, Huntsville, Alabama

Since the completion of the sequencing of the human genome, the JGI has rapidly expanded its scientific goals in several DOE mission-relevant areas. At the JGI-HudsonAlpha, we have kept pace with this rapid expansion of projects with our focus on assessing, assembling, improving and finishing eukaryotic whole genome shotgun (WGS) projects for which the shotgun sequence is generated at the Production Genomic Facility (JGI-PGF). We follow this by combining the draft WGS with genomic resources generated at JGI-HA or in collaborator laboratories (including BAC end sequences, genetic maps and FLcDNA sequences) to produce an improved draft sequence. For eukaryotic genomes important to the DOE mission, we then add further information from directed experiments to produce reference genomic sequences that are publicly available for any scientific researcher. Also, we have continued our program for producing BAC-based finished sequence, both for adding information to JGI genome projects and for small BAC-based sequencing projects proposed through any of the JGI sequencing programs. We have now built our computational expertise in WGS assembly and analysis and perform eukaryotic whole genome shotgun assembly. We have concentrated our assembly development work on large plant genomes and complex fungal and algal genomes. We recently relocated our facility from Stanford University in California to Huntsville, Alabama and are currently in the process of rebuilding our team to in order to have a greater impact on JGI genome projects.

Application of 454 Barcoding to a Difficult Genome

Harris Shapiro[1]* (hjshapiro@lbl.gov), Wilson Phung,[1] Jan-Fang Cheng,[1] Alex Copeland,[1] Raman Koul,[2] Peter Lammers,[2] and Francis Martin[3]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]New Mexico State University, Las Cruces; and [3]INRA-Nancy, France

The whole-genome shotgun (WGS) strategy is the current standard method for sequencing new genomes. It offers many advantages, including relative simplicity of sample preparation and a well-developed variety of assembler packages for putting the pieces together afterwards. However, some genomes are less well suited to this approach than others. In particular, genomes that are highly repetitive and/or polymorphic can be particularly difficult to assemble from WGS data.

For difficult genomes, a fosmid- or BAC-based tiling path presents an alternate sequencing strategy. By pre-binning the assembly into relatively small segments, this strategy circumvents most of the repeat-related issues faced by WGS assembly. In addition, by sequencing only a single haplotype at a time, a tiling path avoids collapsing multiple variants into a single consensus sequence. In the past, the need to separately sequence the large number of clones needed for a tiling path, as well as the need to construct the tiling

path itself, resulted in the WGS approach being favored. However, the development of the ability to sequence large numbers of clones in parallel, using new technology platforms, makes it feasible to reconsider the tiling path technique.

We present the initial results of an experiment performed on a particularly difficult genome, *Glomus intraradices*. This genome shows evidence of being both highly repetitive and polymorphic, making it very challenging to assemble using a WGS approach. As a preliminary step in investigating the feasibility of a tiling path approach, a batch of 36 barcoded fosmids was sequenced using a quarter of an unpaired 454 Titanium run. We describe the procedure used to select the clones and analyze the results. Despite the global complexity of the genome, over one-third of the fosmids (13 of 36) assembled into a single complete contig, while most of the rest assembled into a relatively small number of pieces. As the experiment is ongoing, we describe the next steps that are planned for it.

Functional Relatedness of Energy Producing *Rhodopseudomonas palustris* Genomes

Shaneka Simmons[1,2]* (shaneka.s.simmons@jsums.edu), Hari H.P. Cohly,[1] and Raphael D. Isokpehi[1] (raphael.isokpehi@jsums.edu)

[1]Center for Bioinformatics and Computational Biology, Department of Biology and [2]Environmental Science PhD Program, Jackson State University, Mississippi

We have observed that 82% (538 of 624) completely sequenced prokaryotic genomes contain genes that encode proteins that have the universal stress protein domain (USP, Pfam00582). Furthermore, the following bacterial species: *Geobacter metallireducens, Shewanella oneidensis MR-1, Rhodopsuedomonas palustris, Desulfovibrio vulgaris, Deinococcus radiodurans R1, Nitrosomonas europaea*, and *Clostridium thermocellum*, have been sequenced by the Department of Energy's Joint Genome Institute and are relevant to research on sustainable bioenergy. Among these bacteria, only *Rhodopsuedomonas* had the highest number of sequenced genomes as well as highest number of genes encoding proteins with the USP domain. The universal stress proteins provide biological cells with the ability to respond to environmental stresses such as nutrient starvation, drought, high salinity, extreme temperatures and exposure to toxic chemicals. *Rhodopseudomonas* are rod-shaped, gram-negative, purple nonsulfur, anoxygenic, phototrophic bacteria belonging to the alpha subclass of the Proteobacteria, commonly found in various types of marine environments and soils. These ubiquitous, organisms can grow in both anaerobic and aerobic conditions and have a genetic makeup that allows their genes to be easily removed from their system. *Rhodopseudomonas palustris* are metabolically versatile species that can convert atmospheric carbon dioxide into biomass, recycle aromatic polymers of ligin, produce hydrogen gas for energy production, and fix atmospheric nitrogen. Their ability to adapt and live under various environmental constraints as well as biodegrade pollution to be used as biofuel, make them a model system for research on renewable energy from biological sources. The sequence of strain DX-1 is in production, a strain which is known to produce high power densities that allow it to generate bioelectricity from the biodegration of organic and inorganic waste in low-internal-resistance microbial fuel cells. As of February 6, 2009, there were six finished genome sequences of *R. palustris* strains: BisA53, BisB5, BisB18, CGA009, HaA2 and TIE-1 available in the Joint Genome Institute's Integrated Microbial Genome (IMG) Database. The gene count of these *R. palustris* strains ranged from 4492 to 5377. The objective of this study was to compare the functional relatedness of finished genomes of

R. palustris based on Cluster of Orthologous protein Groups; Protein Families (PFAM), Enzyme and TIGRFAM. Hierarchical Clustering and Correlation Matrix were generated using tools on the IMG database. In the protein family clustering, strains HaA2, CGA009 and TIE-1 clustered on the same branch with CGA009 and TIE-1 on the same node. However, the Enzyme profile CGA009 was placed on a distinct branch from the other five strains. In strain CGA009, 18.1% of the genes were annotated to encode enzymes compared to 9.78% in the TIE-1 strain. We intend to use additional tools on the IMG database to identify specific enzyme catalyzed pathways that are present or enriched in CGA009. Such knowledge may provide insights into strain-specific response and adaptation that could be useful for optimizing engineered strains of *R. palustris* capable of stress tolerance and enhanced biofuel production. Acknowledgements: DHS 2007-ST-104-000007; NSF-EPS-0556308; Research Centers in Minority Institutions (RCMI) – Center for Environmental Health (NIH-NCRR 2G12RR013459-11)

Assessment, Optimization and Applications of 454 FLX Titanium Sequencing Systems

Kanwar Singh (ksingh@lbl.gov), **Zhiying Jean Zhao*** (ZYZhao@lbl.gov), **Natasha Zvenigorodsky*** (NZvenigorodsky@lbl.gov), **Jeff Froula*** (JLFroula@lbl.gov), Len A. Pennacchio, and Feng Chen (fchen@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

Next generation DNA sequencing provides new opportunities to efficiently accomplish a variety of genomic tasks such as the *de novo* assembly of genomes. The newly released 454 FLX Titanium in October 2008 advanced the previous 454 FLX Standard system with nearly double the sequence reads and read lengths. A thorough assessment and optimization of 454 FLX Titanium sequencing system was done by the Technology Development Group at Joint Genome Institute. Currently applications like *de novo whole genome shotgun assembly, transcriptome profiling, pyrotag sequencing and sequence capture technology* are being optimized on the new 454 FLX Titanium systems.

Using *Brachypodium distachyon* to Study the Type II Cell Wall of Grasses

Michael Steinwand[1]* (michael.steinwand@ars.usda.gov), Ludmila Tyler,[2] and John Vogel[1]

[1]USDA ARS Western Regional Research Center, Albany, California; and [2]University of California, Berkeley

With an increased emphasis on renewable sources of energy and the production of sustainable biofuels in current government policies, the search is ongoing for feedstocks to efficiently produce petroleum fuel alternatives. Among the species proposed for use in the production of cellulosic biofuels are several members of the grass family, whose type II cell walls are distinct from the type I cell walls of the model plant *Arabidopsis thaliana* and other dicots. Therefore, there is a need to identify the genes that control the composition and structure of the type II cell wall. Unfortunately, the species proposed as biomass crops (e.g. switchgrass and Miscanthus) are extremely difficult experimental subjects for the types of studies necessary to gain this basic knowledge. Therefore, we have turned to the small model grass *Brachypodium distachyon* (*Brachypodium*).

Brachypodium not only shares the type II walls of other grasses but also possesses the traits necessary to serve as a modern model organism including compact size, rapid generation time, efficient transformation and a small, fully sequenced, genome. We are using a forward genetic approach to identify genes that control cell wall composition and stem structure. As a first step, we optimized conditions for mutagenesis with ethyl methanesulfonate and created a mutant population derived from 8,500 M_1 plants. The M_2 generation is currently being screened for changes in cell wall composition using near-infrared spectroscopy (NIR) and for increased stem density. To date, we have screened 1,000 M_2 plants using NIR and identified 60 putative mutants. These are currently being retested. We have screened 2,000 M_2 plants for stem density and have identified five mutants with stem densities 40% greater than wild type with more retests underway. Increased stem density may translate into more efficient transportation of biomass crops and higher yields. Interestingly, mutants with increased stem density also had an altered NIR profile suggesting that their cell walls are also altered. In addition to the current EMS population, a population of T-DNA insertional mutants will be screened. Future analysis of these mutants will focus on determining what has been changed in the cell wall and if the mutants are easier to ferment. We will use the recently completed whole genome sequence to efficiently clone the genes corresponding to the most promising mutants.

A Community Metagenomic Analysis of the Fungus Gardens from the Leaf-Cutting Ant *Atta colombica*

Garret Suen[1,2]* (gsuen@wisc.edu), Jarrod J. Scott,[1,2] Sandye Adams,[1,2] Frank O. Aylward,[1] Clifton E. Foster,[1,3] Markus Pauly,[1,3] Paul J. Weimer,[1,4] Susannah G. Tringe,[5] Kerrie Barry,[5] Pascal Bouffard,[6] Timothy Harkins,[7] Jolene Osterberger,[7] Steven S. Slater,[1] Timothy J. Donahue,[1,2] and Cameron R. Currie[1,2] (currie@bact.wisc.edu)

[1]Great Lakes Bioenergy Research Center, Madison, Wisconsin; [2]Department of Bacteriology, University of Wisconsin, Madison; [3]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing; [4]U.S. Dairy Forage Research Center, USDA ARS, Madison, Wisconsin; [5]DOE Joint Genome Institute, Walnut Creek, California; [6]454 Life Sciences, Branford, Connecticut; and [7]Roche Diagnostics, Roche Applied Science, Indianapolis, Indiana

For ~50 million years, leaf-cutting ants have been farming fungus for food. The ant-fungus system is one of the most complex symbioses described in nature, and is now known to consist of four mutualists and two pathogens. All members of this symbiosis are located within the fungus gardens of the leaf-cutters, and this system processes hundreds of Kg in dry weight/year of leaf material that is harvested by the ants to grow their mutualistic fungus. The deconstruction of plant biomass within these fungus gardens has particular relevance for bioenergy as a potential source of lignincellulose-degrading microbes. We are investigating the microbial communities in the fungus garden of the leaf-cutting ant *Atta colombica*, which are found in Central and South America. To accomplish this, we employ a combination of culture-dependent and culture-independent methods. We first demonstrate that lignincellulose is degraded by measuring cellulose content as leaf-material passes through the garden. We then performed a 16S metagenome analysis to determine which groups of bacteria are present in the garden. Interestingly, this data suggests that the microbial community is dominated by bacteria that belong to the *Enterobacteriacea*. We then performed a community metagenomic analysis on the fungus gardens to describe the metabolic and physiological potential of the microbial community. Phylogenetic binning of this data confirms that the community is dominated by Enterics,

but revealed the presence of other bacterial lineages that did not appear in our 16S metagenomic analysis. Using this data, we performed directed culturing to obtain representative isolates of these microbes, and are performing functional assays and genome sequencing to confirm their putative roles within the garden. Furthermore, we have performed comparative metagenomics analyses to determine how the carbohydrate metabolism potential of this system compares to that of other sequenced metagenomes. In this way, a clearer picture of how the fungus garden microbial community contributes to the breakdown of lignincellulose can be ascertained, thereby facilitating the identification and characterization of those microbes and lignincellulases that are valuable for bioenergy.

Genomic Resources for the Coral Holobiont–EST Datasets for Two Caribbean Corals and Associated Symbionts

Shinichi Sunagawa[1] (ssunagawa@ucmerced.edu), **Christian Voolstra**[1]* (cvoolstra@ucmerced.edu), Michael DeSalvo,[1] Collin Closek,[1] Erika Lindquist,[2] Jodi Schwarz,[3] Alina Szmant,[4] Mary Alice Coffroth,[5] and Mónica Medina[1]

[1]University of California, Merced; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Vassar College, Poughkeepsie, New York; [4]University of North Carolina, Wilmington; and [5]State University of New York, Buffalo

Funding by the Joint Genome Institute's (JGI) Community Sequencing Program allowed us to generate cDNA libraries and EST datasets (Sanger reads) for two different Caribbean corals, *Acropora palmata* and *Montastraea faveolata*. Additionally, the sequencing of approximately 800,000 ESTs (454 Titanium reads) is underway, which will include two of their dinoflagellate symbionts, *Symbiodinium* sp. CassKB8 and *Symbiodinium* sp. Mf1.05b. Some of these resources have been used to print cDNA microarrays to profile gene expression patterns related to the coral's life cycle—onset, maintenance, and breakdown (bleaching) of symbiosis. Currently, we are expanding our coral microarrays and will extend our platforms to study the coral holobiont from both the host's and symbiont's perspectives. In addition to transcriptome profiling, one of our goals is to further identify and analyze coral- and *Symbiodinium*-specific genes via molecular evolution (dN/dS) analyses. These genes are prime candidates to generate lineage-specific phenotypic traits. A publicly available database generated by our lab hosts the assembled and annotated datasets and will continue to incorporate genomic resources for symbiotic cnidarians.

Elucidating QTL for Yield and Saccharification in Bioenergy Poplar

Gail Taylor[1,2]* (g.taylor@soton.ac.uk), Patrick D. Stephenson,[1] Matt D. Nelson,[1] and Anja Bus[1,2]

[1]University of Southampton, United Kingdom; and [2]The Porter Alliance, Imperial College, London

Our aim is to elucidate genes determining yield and cell wall disassembly in bioenergy poplar grown as short rotation coppice. Following two coppice cycles where a mapping population of *Populus trichocarpa* x *Populus deltoides* was grown for six years in a fully replicated randomised trial, we harvested woody stems and subjected them to saccharification using two different saccharification protocols. At the same time, we have

studied this population intensely and identified the ideotype for bioenergy yield with early diagnostic traits for yield defined (Rae et al., 2004). More recently we have confirmed five biomass yield QTL hotspots (Rae et al., 2009) and report the detailed investigation of one of these using a genetical genomics approach. From positional SNPs genotyping and identification of 'high' and 'low' biomass extremes, we have narrowed the genes determining this QTL and out latest findings will be reported. One QTL in particular accounts for more than 10% of phenotypic variability. Similarly QTL for saccharification have also been identified and one targeted for on-going further elucidation.

In a complementary approach we are using a wide association population of *Populus nigra* collected from contrasting sites across Europe and this is being used to elucidate yield-gene and saccharification-gene associations.

- Rae AM, Robinson KM, Street NR and Taylor G (2004). Morphological and physiological traits influencing biomass productivity in short rotation coppice poplar. *Can J For Res* 34: 1488-1498
- Rae, AM Street NR, Harris N and Taylor G (2009). Five QTL hotspots for yield in short rotation bioenergy poplar: The Poplar Biomass Loci, *BMC Plant Biology* (In Press).

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, Albany, California; and [2]Samuel Roberts Noble Foundation, Inc., Forage Improvement Division, Ardmore, Oklahoma

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 a *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 betaglucan 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 3.18 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.

The JGI Metagenome Program

Susannah G. Tringe* (sgtringe@lbl.gov), Martin Allgaier, Kerrie Barry, and Philip Hugenholtz

DOE Joint Genome Institute, Walnut Creek, California

The JGI has been a leader in metagenomic research and has a large and growing metagenome program encompassing a variety of technical and analytic approaches. We are currently in the process of transferring the bulk of our standard metagenome sequencing from Sanger to 454-titanium sequencing technology, and this poster will report on the impact the change in technologies has had on downstream assembly and analysis. We are also developing laboratory protocols and analysis pipelines for metatranscriptomics, 16S pyrotag analysis and pooled fosmid sequencing as complements to our metagenome sequencing capabilities. Progress in these areas will be discussed along with scientific examples from our metagenome portfolio.

Users' Annotation of the *Chlamydomonas* Genome: A Tale of Two Cultures

Olivier Vallon[1]* (ovallon@ibpc.fr) and Sabeeha Merchant[2]

[1]UMR 7141 CNRS/UMPC, Institut de Biologie Physico-Chimique, Paris, France; and [2]Department of Chemistry and Biochemistry, University of California, Los Angeles

After the initial jamboree presenting the automatic annotation of the *Chlamydomonas* genome (draft version 2.0, December 2003), the need emerged for a coordination of the "manual curation" effort. We have organized the community towards this task, enlisting over a hundred experts as annotators with writing access to the database. Over 5 years, these annotators have worked on 5,767 loci, entering gene names, deflines and descriptions that were subsequently uploaded to Genbank together with the gene sequences. While the contribution of each annotator varied broadly, both in quality and in quantity, the overall impact on the accuracy and usefulness of the annotation was enormous. Expert annotation helped filtering out "non-genes", enhanced the transcript description at many loci (either

by choosing a better model from those made available on the browser, or by modifying them or even by creating new models de novo), and enhanced the functional annotation with gene names, detailed description of function, literature references etc… The team included *Chlamydomonas* experts, as well as scientist working with other organisms but willing to apply their knowledge and specialized annotation tools to the task of annotating the genome of a prominent experimental model. A wealth of specialized annotation papers have been published, before and after the release of the "genome paper" in Science. Many annotators have gone one step further and introduced *Chlamydomonas* in their laboratory.

Now that version 4 of the genome has been released, the time has come to analyze the impact that this effort has had on the quality of the annotation and discuss the future of this organisation scheme. Can the feeble human eye outperform a trained automatic annotation algorithm, in terms of structural or functional annotation? How can we deal with human errors and occasional sloppiness? What with the many genes that have not been manually curated? What happens when a new version of the genome is released, and annotation needs to be mapped forward? Can such a working model be implemented for organisms that are used in fewer labs? How can the interface between JGI and the community be improved to facilitate the work of both?

Rapid Marker Generation in Soybean Using Next-Generation Sequencing

Kranthi Varala[1]* (kvarala2@uiuc.edu), Ying Li,[1] Kankshita Swaminathan,[2] and Matthew Hudson[1]

[1]Department of Crop Sciences and [2]Energy Biosciences Institute, University of Illinois, Urbana-Champaign

North American soybean is known to have gone through genetic bottlenecks during its introduction from the east-Asian accessions. Numerous desirable traits, such as biotic and abiotic stress resistance, are often found in exotic accessions or wild soybean. Soybean breeders have relied on a mixture of RAPD, SSR and few SNP markers to guide the integration of these traits into the elite soybean cultivars. Zhu et al. (2003) estimated the sequence variation in cultivated soybean as 1 SNP/kilobase, based on a 76kbp region from 25 accessions. SNP markers allow rapid genotyping with simple PCR techniques in high throughput. Marker availability, especially SNP markers has been limited in the past and generating new markers for rare accessions has proved time consuming. With the availability of the chromosome-scale assembly of the Soybean *var*. Willams 82 genome and next-generation sequencing technologies that are geared towards rapid SNP detection, the availability of SNP information is projected to increase rapidly. In this study we sequenced a reduced representation library of 4 soybean accessions of interest, Dowling, Dwight, Komata and Rpp1, using the Illumina 1G platform. Williams 82 was sequenced as control for the experiment. Nuclear DNA from these accessions was extracted and digested with MseI to give short fragments anchored at the recognition sequence. The enzyme choice was aimed to selectively cut within non-repetitive sequences with a greater frequency, as derived from the survey sequencing results by Swaminathan et.al (2007). Approximately 9 million reads were obtained from each accession of which ~2.7 million aligned in a unique location. All alignments were performed with maq. Each of the accessions sequenced identified significantly higher SNPs than the control Williams data. SNPs ranged from 3500-15000 and correlate positively with the estimated genetic distance between the lines. The Komata accession was further sequenced using a WGS approach with paired and unpaired runs of Illimuna 1G. This approach greatly increased the SNP

density, albeit with reduced confidence in the SNP calls. A small number of the predicted SNPs from both approaches were confirmed using PCR amplification and Sanger sequencing in the lab.

Assembly and Annotation of the *Brachypodium distachyon* Genome

John Vogel[1]* (john.vogel@ars.usda.gov), Klaus Mayer,[2] Daniel Rokhsar,[3] Jeremy Schmutz,[4] Todd Mockler,[5] Naxin Huo,[1] Yong Gu,[1] David Garvin,[6] and Michael Bevan[7]

[1]USDA ARS Western Regional Research Center, Albany, California; [2]MIPS, Munich, Germany; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Hudson Alpha Institute of Biotechnology, Huntsville, Alabama; [5]Oregon State University, Corvallis; [6]USDA ARS Plant Science Research Unit, University of Minnesota, St. Paul; and [7]John Innes Centre, Norwich, United Kingdom

Brachypodium distachyon (*Brachypodium*) is rapidly emerging as a model system to study questions unique to the grasses. This emergence is coincident with an increased need for basic research in grass biology to develop perennial grasses as a source of renewable fuel. The list of genomic resources available to *Brachypodium* researchers is increasing exponentially. Resources we have developed include: a highly efficient transformation system, a high density SNP-based genetic map, a physical map, BAC libraries and BAC end sequences, EST sequences, and mutagenesis protocols. In addition, the generation of a sequence-indexed insertional mutant population is underway with >3,000 mutants generated to date. We recently completed the sequencing and preliminary analysis of the complete genome. The final assembly and annotation was of extremely high quality and promises to be a powerful research tool. When taken together, these resources enable researchers to utilize *Brachypodium* for a wide array of experimental approaches, including those that require complete genome sequence. An overview of the *Brachypodium* genome project will be presented.

SNPs of Information: Inferring Evolutionary History in *Coccidioides*

Emily Whiston[1]* (whiston@berkeley.edu), Daniel E. Neafsey,[2] Chiung-Yu Hung,[3] Jason E. Stajich,[1] Thomas J. Sharpton,[1] Cody McMahan,[3] Gary T. Cole,[3] and John W. Taylor[1]

[1]Department of Plant and Microbial Biology, University of California, Berkeley; [2]Broad Institute, Massachusetts Institute of Technology, Cambridge; and [3]Department of Biology, University of Texas, San Antonio

Coccidioides spp., the causative agent of Coccidioidomycosis, is a dimorphic fungus, with a saprobic hyphal phase and a pathogenic spherule phase. We have previously shown that there are two species of *Coccidioides*: *C. immitis* (found in California and Mexico) and *C. posadasii* (found in Arizona, Texas, Mexico and South America). Recently, 4 strains of *C. immitis* and 10 strains of *C. posadasii* have been sequenced. Using a set of high-quality SNPs in these genomes, we have investigated effective population size, comparative SNP rates across different genomic features, patterns of positive selection in coding regions, and levels of conservation among vaccine candidates. We see that *C. posadasii* has a 2-fold larger effective population compared to *C. immitis*, but that *C. immitis* has more genes undergoing positive selection. Additionally, we have identified a set of conserved potential vaccine candidates. Using these data, we can make further hypotheses about the

evolutionary history of *Coccidioides spp*. and inform the development of vaccine candidates.

The Lipid Body Proteome of the Marine Haptophyte Alga *Emiliania huxleyi*: Potential for Biofuels

Gordon Wolfe* (gwolfe2@csuchico.edu)

Department of Biological Sciences, California State University, Chico

Emiliania huxleyi and some related haptophyte algae produce as neutral lipids a set of Poly-*trans*-Unsaturated Long-Chain (C_{37-39}) Alkenones, alkenoates, and alkenes (abbreviated PULCA). These biomarkers are widely used for paleothermometry, but the biosynthesis and cellular location of these unique lipids remain largely unknown. Eltgroth et al. (2005) previously showed that these taxa, like many other algae, package their neutral lipid into cytoplasmic vesicles or lipid bodies (LBs). Here, I present a proteomics analysis of isolated lipid body proteins, based on *E. huxleyi* CCMP 1516, sequenced recently by DOE-JGI; this is the first proteomic study of this important alga.

I developed a method for fractionation and separation of PULCA-rich fractions from *Emiliania*, as well as the sister taxa *Isochrysis*, and analyzed their low-abundance protein fractions. Following trypsin digestion, mass spectrometric analysis revealed 509 unique protein IDs from the v. 1 Ehux genome. Of these, 61 were highly abundant in the LB fraction, of which 66% were unique to this fraction, while the rest were highly enriched. Roughly 70% had EST support, and less than 10% could not be ascribed known function. Proteins included those for structure (actin, PAP-fibrillin), acyl and prenyl lipid biosynthesis, β-oxidation, as well as proteins for signaling, translation and trafficking, and chaperones. The overall LB proteome strongly resembles those from other photosynthetic organisms, especially *Chlamyodomonas*, and suggests tight coupling of LBs with endomembranes, peroxisomes and mitochondria, and the secretory system, as with other taxa. Comparison with *Arabidopsis thaliania* acyl lipid pathways strongly suggests fatty acids as the biosynthetic precursors. Interestingly, no fatty acid desaturases were found in the LB fraction, despite their abundance in the genome. Rather, some novel zeta-carotene-desaturase-like proteins are abundant, and may be involved in modification of these unusual acyl lipids.

Surprisingly, in *Emiliania* but not *Isochrysis*, netural lipid localizes to the cell wall as well, and proteins associated with secretion are evident in LB proteome fractions. When cells were fed bicarbonate in conjunction with N or P limitation, I obtained massive stimulation of both PULCA and LBs, and total LB protein increased, but so did proteolysis. These conditions also led to massive stimulation of calcification in some *Emiliania* strains, already well known to be due to cell arrest in G1. Therefore, it appears that cells can be stimulated to shunt energy and reducing power into both hydrocarbon and calcite production under nutrient-limited bicarbonate stimulation, and the association of neutral lipid with cell walls and lithification suggests a possible mechanism to export lipids for harvest in continuous culture. PULCA are potentially preferable to triacylglycerides (TAG) as biofuels, as these poly-*trans* unsaturated pure hydrocarbons have no glycerol, and are also much more resistant to photooxidation. Along with the ability to grow these algae in a wide variety of saline waters, these features suggest a possible source of biofuels that deserves further study.

Genome Survey Sequencing of Shiitake Mushroom *Lentinula edodes*: Comparing with Sequenced Basidiomycete Genomes

M.C. Wong* (manchun3@hotmail.com), C.H. Au, Winnie W.Y. Chum, P.Y. Yip, Iris S.W. Kwok, Patrick T.W. Law, and H.S. Kwan

The Chinese University of Hong Kong, HKSAR, PR China

Shiitake mushroom *Lentinula edodes* is a popular edible mushroom of high economic values. However, the molecular mechanisms involved in its growth and development is not fully understood. Next-generation sequencing was employed for genome sequencing to elucidate the genomic architecture of *L. edodes*. Bioinformatics analysis was carried out to compare the protein-coding genes of *L. edodes* with that of five other sequenced basidiomycetes. A pilot-scale genome survey sequencing (GSS) of dikaryotic *L. edodes* L54 was performed using Roche Genome Sequencer GS-20, resulting in 353,030 reads (total length 35.7Mbp). Then, a scaled-up genome shotgun sequencing of monokaryotic strain L54-A was carried out using Roche GS-FLX system, generating 506,370 reads (total length 116.8Mbp). Together with the 12,000 unpublished in-house cDNA contigs of *L. edodes* and the 12,210 publicly available *L. edodes* expressed sequence tag (EST) sequences in NCBI dbEST, all *L. edodes* sequences were searched against protein-coding genes of five sequenced basidiomycetes, namely *Coprinopsis cinerea, Laccaria bicolor, Phanerochaete chrysosporium, Cryptococcus neoformans* and *Ustilago maydis*. From 54% (*L. bicolor*) to 78% (*P. chrysosporium*) of protein-coding gene homologs are present in *L. edodes* sequences.

Assembling the Marine Metagenome, One Cell at a Time

Tanja Woyke* (twoyke@lbl.gov), Gary Xie, Alex Copeland, José M. González, Cliff Han, Hajnalka Kiss, Jimmy Saw, Pavel Senin, Chi Yang, Jan-Fang Cheng, Michael E. Sieracki, and Ramunas Stepanauskas

DOE Joint Genome Institute, Walnut Creek, California

Determining the genetic makeup of predominant microbial taxa with specific metabolic capabilities remains one the major challenges in microbial ecology and bioprospecting, due to the limitations of current cell culturing and metagenomic methods. The complexity of microbial communities and intraspecies variations poses obstacles for the assembly of individual genomes from metagenomic shotgun libraries. Here we report the use of single cell genomics to access the genome of two uncultured proteorhodopsin-encoding flavobacteria from Gulf of Maine. High throughput fluorescence-activated sorting of single cells, whole genome amplification via multiple displacement amplification and PCR-screening enabled shotgun sequencing of these single amplified genomes (SAGs) yielding 1.9 Mbp (17 contigs) and 1.5 Mbp (21 contigs) draft genomes for the two flavobacteria. In contrast to cultured strains, the two uncultured flavobacteria genomes were excellent Global Ocean Sampling (GOS) metagenome fragment recruiters, demonstrating their numerical significance in the ocean. Annotation revealed genome streamlining and diversified energy sources of the two uncultured microorganisms, including biopolymer degradation, proteorhodopsin photometabolism, and H_2 oxidation. These features may be indicative for specific adaptations to the marine environment and for the absence of related microorganisms in cultures.

Phylogenetic Diversity Contributions of the Genomic Encyclopedia for Bacteria and Archaea Pilot Project

Dongying Wu[1]* (dygwu@ucdavis.edu), Phil Hugenholtz,[2] Nikos Kyrpides,[2] and Jonathan A. Eisen[1]

[1]University of California Genome Center, Davis; and [2]DOE Joint Genome Institute, Walnut Creek, California

In order to fill the phylogenetic gaps of current bacterial and archaeal genome sequences, we started the Genomic Encyclopedia for Bacteria and Archaea pilot project (GEBA) in which the selection of microbial organisms for sequencing were guided by the phylogenetic novelty of organisms relative to those for which complete genomes have been determined or are in progress. We have released the sequences of 53 bacterial and 3 archaeal genomes from the pilot project. In this study, we built a maximum likelihood "genome" tree for currently available bacteria genomes including those from the GEBA bacterial genomes. The tree was built upon the concatenated protein alignments of 31 phylogenetic markers and was used for comparing the phylogenetic diversity contributions of GEBA bacterial organisms with previous available genomes. We've built gene families for GEBA genomes, as wells as genomes selected from taxonomic groups at different phylogenetic levels, to exam the impact of phylogenetic diversity on the rate of recovery of novel gene families. We've found that the phylogenetic contributions of the GEBA pilot project are significantly larger than randomly selected bacterial genomes, and phylogenetic diversity is one of the key factors to affect the rate of novel gene family discoveries. Using actinobacterial gene families as an example, we detailed the contributions of GEBA genomes in terms of gene families characteristic changes and novel gene families discoveries. Given the significant phylogenetic diversity contribution of the GEBA pilot project, we estimate our future genome selection based on a 16S tree built from the greengenes project: Current genome sequences only cover 2.2% of the PD if we considered all the bacteria and archaea in the non-redundant Greengenes ss-rRNA dataset; to cover 50% of the phylogenetic diversity, we have to sequences 9218 more genomes.

Sequencing of T-DNA Flanking Regions in *Brachypodium distachyon*

Jiajie Wu[1,2]* (Jiajie.wu@ars.usda.gov), Jennifer Bragg,[2] John Vogel,[2] Gerard Lazo,[2] Olin Anderson,[2] and Yong Gu[2]

[1]University of California, Davis; and [2]USDA ARS, Western Regional Research Center, Albany, California

Brachypodium distachyon (*Brachypodium*) is being developed as a truly tractable model system for the grasses. It is also an ideal model for studying the basic biology underlying the traits that control the utility of grasses as energy crops. Several important genomic resources have been developed in *Brachypodium*, including the complete genome sequence, ESTs, SNP markers, a high-density linkage map and germplasm resources. To facilitate functional genomic research, we have developed an extremely efficient *Agrobacterium*-mediated transformation method and have begun generating a population of T-DNA mutants. Here we describe the sequencing of T-DNA flanking regions in this population.

An AGP Pipeline for Maize Genome Sequencing Project

Jianwei Zhang[1]* (jzhang@cals.arizona.edu), Fusheng Wei,[1] Maize Genome Sequencing Consortium,[1,2,3,4] and Rod A. Wing[1]

[1]Arizona Genomics Institute, Department of Plant Sciences, University of Arizona, Tucson; [2]Genome Sequencing Center,Washington University School of Medicine, St. Louis, Missouri; [3]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; and [4]Iowa State University, Ames

Maize is an important food source for animals and man, as well as a premier model biological system. Research on maize can also be beneficial for better understanding its relatives such as sorghum, wheat, rice. The Maize Genome Sequencing Project used a BAC-by-BAC strategy to sequence the genome and provided a high quality reference sequence in low copy regions. Because over 80% of the genome is repetitive, the resulting sequences in each BAC contain multiple sequence contigs (11 on average per clone) that are neither ordered nor orientated. It is therefore quite difficult to build BAC-based maize pseudochromosomes. We setup an "A Gold Path (AGP)" pipeline to generate these pseudomolecules based on the integrated genetic and physical map built with markers and sequences. This pipeline is a semiautomatic web-based system for data management and analysis. First, we collected all available data including sequences of BACs, BAC ends and markers from both our sequencing project and outside projects. Then a serial set of comparisons between neighboring BAC sequences and/or BES were undertaken by running BLAST searches. These results helped us to locate the left and/or right ends of each BAC on the adjacent BACs, as well as the overlapping regions between two adjoining clones. Thus we can easily determine the precise order and orientation of some pieces in a BAC sequence and remove the overlap from the final pseudomolecule according to the sequence alignment. All processing information was saved into our database for final output of an AGP file. Additionally most data can be browsed or modified with a user friendly graphical interface. At present, a total of 16909 clones assigned to FPC 422 contigs were collected in the pipeline. Lastly we obtained about 2060Mb of pseudomolecule sequence, which covers ~97% of the 2120-Mb physical map. Nearly 850-Mb of redundant or overlapping sequences was removed in the Phase I pipeline. Further sequence ordering and orientating is being conducted utilizing maize-sorghum and maize-rice syntenic relationships in the Phase II analysis. The data will soon be available at www.genome.arizona.edu, and uploaded in the project web site www.maizesequence.org.

Development of Resources and Tools for Switchgrass Functional Genomics

Ji-Yi Zhang[1,4]* (jzhang@noble.org), Yi-Ching Lee,[1,4] Ivone Torres-Jerez,[1] Eric Worley,[1,4] Jiading Yang,[1,4] Mingyi Wang,[1] Ji He,[1] Yuhong Tang,[1] Christa Pennacchio,[3] Erika Lindquist,[3] Malay Saha,[2] and Michael Udvardi[1,4]

[1]Plant Biology Division and [2]Forage Improvement Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma; [3]DOE Joint Genome Institute, Walnut Creek, California; and [4]DOE Bioenergy Science Center

Switchgrass (*Panicum virgatum* L.) is an outbreeding, fast-growing perennial C_4 grass species native to North America, which is used as forage on the Great Plains of the USA and has the potential to become a major source of biomass for bio-fuel production. To

realize this potential, breeding efforts are underway to improve wild germplasm and forage type switchgrass for bioenergy uses. Development of genomics resources will aid the breeding enterprise, via marker development, and enable forays into molecular and systems biology of this species.

Alamo AP13, a lowland genotype, and Summer VS16, an upland genotype, are the parents of a mapping population that is currently being characterized in the field for traits related to biomass/biofuel production. We have clonally-propagated these two genotypes, via vegetative cuttings, to generate a sufficient number of genetically-identical individuals for RNA isolation, cDNA synthesis and library construction, and sequencing. EST sequencing of these two genotypes will facilitate SNP discovery for genetic mapping of target traits by switchgrass breeders.

To date, we have generated 1.4 million ESTs from Alamo AP13 cDNA libraries of roots and shoots harvested at different stages of development. *De novo* assembly has been conducted using the MALIGN and CAP3 programs. 31,440 tentative consesus sequences (TCs) and 462,279 unigene sequences from shoots, and 27,616 TCs and 483,029 unigenes from roots have been obtained. Likewise, 1.5 million ESTs have been generated from multiple cDNA libraries of Summer VS16. Assembly of sequences into TCs has been conducted for each library and super-clustering across libraries is underway.

Two full-length cDNA libraries of Alamo AP13 plant are currently being constructed and 50,000 clones will be sequenced from both ends. This will enable further clustering of AP13 cDNA sequences and provide verified full-length clones for functional analysis of proteins in the future. After completion of EST/cDNA sequencing of both genotypes, sequence alignment programs will be used to identify putative SNPs for genetic mapping in populations derived from the two parents.

A switchgrass sequence database is now available at: http://bioinfo4.noble.org/switchgrass/.

Emulsion PCR Amplification of Nanogram-Scale cDNA For Transcriptome Sequencing

Tao Zhang[1]* (tzhang3@lbl.gov), Mei Wang,[1] Christopher Villalta,[2] Edward Kirton,[1] Erika Lindquist,[1] Andrea Aerts,[1] Igor Grigoriev,[1] John Taylor,[2] and Len A. Pennacchio[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; and [2]Department of Plant and Microbial Biology, University of California, Berkeley

Sample preparation for next generation sequencing platforms typically requires microgram levels of starting material for DNA library construction. However, such DNA quantities are in many cases rate limiting such as for transcriptome sequencing of microbes growing on nutrient poor wood substrates. Previously, we have developed an emulsion PCR (emPCR) method for linear amplification and sequencing of low-DNA-content samples. This method has been successfully applied to sequence nanogram and even sub-nanogram quantities of ancient DNA (Blow MJ, et al. (2008) Genome Research. 18, 1347-53) and ChIP DNA (Visel A, et al. (2009) Nature. 457, 854-859), by using next generation sequencers. To apply this method to sequence transcriptomes, we have ligated 454 linkers to 1 nanogram of cDNA and constructed an emulsion amplified library. In comparison, we also made a control library from 1 microgram of cDNA, using the standard 454 protocol. Upon expression profile analysis, we found that these two libraries are highly correlated, suggesting that the emPCR amplification process does not introduce major biases. Encouraged by this result, we have isolated ~1 nanogram of cDNA from *Neurospora*

discreta colonies growing on *Populus trichocarpa* stem tissues, and are in process of emPCR amplification and sequencing of this sample. We hope to identify genes and pathways in *Neurospora discreta* important for degradation of cellulosic biomass through transcriptome sequencing. We believe that emPCR is a useful tool to make broad range of low-DNA-content samples accessible to next generation analysis.

Fungal Intron Evolution: Why a Small Genome Can Have Many Introns

Kemin Zhou* (kzhou@lbl.gov), Alan Kuo, Afaf Salamov, and Igor Grigoriev

DOE Joint Genome Institute, Walnut Creek, California

Previous phylogenetic intron studies have focused on individual genes and intron locations. The best whole genome studies used only 30 genes. Here we used a statistical approach instead. We analyzed exon number and intron length from whole genomes of seven Basidiomycota, seven Ascomycota, one Zygomycota, and one Chytridiomycota species. A whole genome phylogenetic tree was constructed using 288 proteins which agree with previous publications. Our exon number analysis indicates that the ancestor of fungi have large numbers of exons (from 5.8 to 7.2). *Sporobolomyces roseus* is most related to *Ustilago maydis* based on our phylogenetic tree. Both have small genomes but they differ dramatically in the number of exons per gene (*S. roseus* with an average 7.2 and *U. maydis* 1.7). The large number of exons per gene in *S. roseus* correlates with lack of reverse transcriptase (RT) foot-prints in the genome and lack of symptoms of RT-mediated intron loss. We also found that the number of genes in each genome is directly correlated with the genome size (slope = 4167 nucleotides per gene). We identified a clear difference between Ascomycota and other fungal phyla in the pattern of intron loss with the exception of *U. maydis*, whose pattern resembles that of the Ascomycota yeast *Pichia stipitis*. We also identified a phylogenetically divergent effect of RT on exon numbers per gene. RT correlates with intron loss of Ascomycota but correlates with intron gain in Basidiomycota. In addition we identified RT as a major contributing factor to fungal genome size. Finally we found that the number of exons per gene within the same genome differ between genes of different degrees of conservation, and there is a direct correlation between the number of exons from genes conserved in all 16 genomes and the difference between between this and that of genes that are specific to each species, with a coefficient of 0.5 to 0.54. So we call this The Half Difference Rule.

Attendees

Current as of March 12, 2009

Birte Abt
DSMZ- German Collection of Microorganisms and Cell Cultures Ltd.
bab08@dsmz.de

Catherine Adam
DOE Joint Genome Institute
adam1@llnl.gov

Andrea Aerts
DOE Joint Genome Institute
alaerts@lbl.gov

Musahid Ahmed
Chemical Sciences
mahmed@lbl.gov

Dag Ahren
Microbial Ecology
dag.ahren@mbioekol.lu.se

Nickolai Alexandrov
Ceres, Inc.
hdungca@ceres-inc.com

Manuel Alfaro
Public University of Navarre
manuel.alfaro@unavarra.es

Ed Allen
DOE Joint Genome Institute
eaallen@lbl.gov

Martin Allgaier
Joint BioEnergy Institute
mallgaier@lbl.gov

Gary Andersen
Lawrence Berkeley National Lab
glandersen@lbl.gov

Iain Anderson
DOE Joint Genome Institute
ijanderson@lbl.gov

Amy Anderton
USDA, WRRC
amy.anderton@ars.usda.gov

William Andregg
Halcyon Molecular
wandregg@gmail.com

Dion Antonopoulos
Argonne National Laboratory
dion@anl.gov

Chun Hang Au
The Chinese University of Hong Kong
tommyau@cuhk.edu.hk

Seth Axen
DOE Joint Genome Institute
saxen@lbl.gov

Gyorgy Babnigg
MCSG, Argonne National Laboratory
gbabnigg@anl.gov

Scott Baker
Pacific Northwest National Laboratory
scott.baker@pnl.gov

Venkatesh Balakrishnan
University of Wisconsin
vbalakrishn2@wisc.edu

Elizabeth Barker
University of Regina, SK, Canada
barker1e@uregina.ca

Adam Barry
DOE Joint Genome Institute
abarry@lbl.gov

Kerrie Barry
DOE Joint Genome Institute
kwbarry@lbl.gov

Laura Bartley
University of California, Davis
lebartley@ucdavis.edu

Mary Bateson
Montana State University
mbateson@montana.edu

Diane Bauer
DOE Joint Genome Institute
bauer20@llnl.gov

Bonnie Baxter
Westminster College
bbaxter@westminstercollege.edu

Laura Beer
Colorado School of Mines
lbeer@mines.edu

Sadia Bekal
University of Illinois
sbekal@uiuc.edu

Harry Beller
Lawrence Berkeley National Lab
hrbeller@lbl.gov

John Benemann
Benemann Associates
jbenemann@aol.com

Gry Mine Berg
Stanford University
mineberg@stanford.edu

Stephanie Bernard
Lawrence Berkeley National Lab
smbernard@lbl.gov

Christophe Billette
INRA
billette@bordeau.inra.fr

Emanuele Biondi
University of Florence
emanuele.biondi@unifi.it

Matthew Blow
DOE Joint Genome Institute
mjblow@lbl.gov

Igor Bogorad
University of California, Los Angeles
igor.bogorad@gmail.com

Harvey Bolton
PNNL
harvey.bolton@pnl.gov

Jeffrey Boore
Genome Project Solutions and UC Berkeley
jlboore@genomeprojectsolutions.com

Jennifer Bragg
USDA, ARS, WRRC
jennifer.bragg@ars.usda.gov

Lambert Brau
Murdoch University
lbrau@murdoch.edu.au

Susanna Braus-Stromeyer
Georg-August-University Goettingen
sbraus@gwdg.de

Susan Brawley
University of Maine
brawley@maine.edu

Thomas Brettin
DOE Joint Genome Institute
tsbrettin@lbl.gov

Jim Bristow
DOE Joint Genome Institute
jbristow@lbl.gov

Leticia Britos
Stanford Univ Sch of Medicine
britos@stanford.edu

Eoin Brodie
Lawrence Berkeley National Lab
elbrodie@lbl.gov

Attendees

Pamela Brown
Indiana University
pjbonner@indiana.edu

Shane Brubaker
LS9, Inc.
brubaker@ls9.com

David Bruce
DOE Joint Genome Institute
dbruce@lanl.gov

Thomas Bruns
University of California, Berkeley
pogon@berkeley.edu

Elizabeth Buchen
Freelance
lizziebuchen@gmail.com

Kathryne Byrne-Bailey
University of California, Berkeley
kbyrne@nature.berkeley.edu

Chenier Caoile
HudsonAlpha Genome Sequencing Center
ccaoile@hudsonalpha.org

David Casero
University of California, Los Angeles
dcasero@ucla.edu

Jamie Cate
EBI-Berkeley
jcate@lbl.gov

Jean Challacombe
JGI-LANL
jchalla@lanl.gov

Mike Challen
University of Warwick
mike.challen@warwick.ac.uk

Michele Champagne
Kuehnle AgroSystems Inc.
michele@kashawaii.com

Patricia Chan
University of California, Santa Cruz
pchan@soe.ucsc.edu

Yun-juan Chang
BSD
yjs@ornl.gov

Erika Check Hayden
Nature
e.check@naturesf.com

Fanqing Chen
Lawrence Berkeley National Lab
f_chen@lbl.gov

Feng Chen
DOE Joint Genome Institute
fchen@lbl.gov

Jan-Fang Cheng
DOE Joint Genome Institute
jfcheng@lbl.gov

Olga Chertkov
Los Alamos National Lab
ochrtkv@lanl.gov

Swapnil Chhabra
LBL/JBEI
srchhabra@lbl.gov

Mansi Chovatia
DOE Joint Genome Institute
mrchovatia@lbl.gov

Julianna Chow
DOE Joint Genome Institute
jchow@lbl.gov

Katy Christiansen
Joint BioEnergy Institute
kmchristiansen@lbl.gov

George Church
Harvard Medical School
g1m1c1@receptor.med.harvard.edu

Frank Collart
Argonne National Lab
fcollart@anl.gov

Alex Copeland
DOE Joint Genome Institute
accopeland@lbl.gov

Jean-Philippe Coppé
Dynamic Throughput Inc.
tribur@gmail.com

Robert Cottingham
Oak Ridge National Laboratory
cottinghamrw@ornl.gov

Michael Cox
University of Wisconsin-Madison
pfeffer@biochem.wisc.edu

April Cunningham
DOE Joint Genome Institute
acunningham@lbl.gov

Cameron Currie
University of Wisconsin-Madison
currie@bact.wisc.edu

Patrik D'haeseleer
LLNL, JBEI
patrikd@gmail.com

Jeff Dangl
Univ. of North Carolina
dangl@email.unc.edu

Richard Danielson
BioVir Laboratories, Inc.
red@biovir.com

Chris Daum
DOE Joint Genome Institute
daum1@llnl.gov

Dhwani Desai
Leibniz Institut fur Meereswissenschaften (IFM-GEOMAR)
ddesai@ifm-geomar.de

Shweta Deshpande
DOE Joint Genome Institute
sdeshpande@lbl.gov

Chris Detter
Los Alamos National Lab
cdetter@lanl.gov

Samuel Deutsch
DOE Joint Genome Institute
sdeutsch@lbl.gov

Maria Dominguez-B
University of Puerto Rico
mgdbello@uprr.pr

Victor Dorsett
DOE Joint Genome Institute
vtdorsett@lbl.gov

Daniel Drell
US Department of Energy
daniel.drell@science.doe.gov

Inna Dubchak
JGI, LBNL
ildubchak@lbl.gov

Evi Dube
DOE Joint Genome Institute
edube@lbl.gov

Erin Dunwell
JGI (LLNL)
dunwell2@llnl.gov

Ashlee Earl
Harvard Medical School
ashlee_earl@hms.harvard.edu

Daniel Eastwood
University of Warwick
daniel.eastwood@warwick.ac.uk

Joseph Ecker
The Salk Institute for Biological Studies
ecker@salk.edu

Rob Egan
DOE Joint Genome Institute
rsegan@lbl.gov

Levente Egry
Applied Biosystems
levente.egry@appliedbiosystems.com

Jonathan Eisen
UC Davis Genome Center
jaeisen@ucdavis.edu

Tedd Elich
GrassRoots Biotechnology
tedd.elich@grassrootsbio.com

Christopher Ellison
Plant & Microbial Biology, UC Berkeley
cellison@berkeley.edu

Anna Engelbrektson
DOE Joint Genome Institute
aengelbrektson@lbl.gov

Joseph Fass
University of California, Davis
jnfass@ucdavis.edu

Joni Fazo
DOE Joint Genome Institute
jbfazo@lbl.gov

Heidi Feiler
Lawrence Berkeley National Lab
hsfeiler@lbl.gov

Marsha Fenner
DOE Joint Genome Institute
mwfenner@lbl.gov

James Fey
DOE Joint Genome Institute
jfey@lbl.gov

Klaus Fiebig
Ontario Genomics Institute
kfiebig@ontariogenomics.ca

Antonella Fioravanti
University of Florence
antofiora@gmail.com

Nathan Fisher
USAMRIID
nathanfisher1@gmail.com

Christopher Francis
Stanford University
caf@stanford.edu

Tracey Allen Freitas
University of Hawaii, Manoa
tracey.freitas@gmail.com

Susan Fuerstenberg
Genome Project Solutions
sifuerst@genomeprojectsolutions.com

Craig Furman
DOE Joint Genome Institute
cfurman@lbl.gov

Steven Garan
Lawrence Berkeley National Lab
sgaran@lbl.gov

Hector Garcia Martin
Joint BioEnergy Institute
hgmartin@lbl.gov

Scott Geib
Penn State University
smg283@psu.edu

Terry Gentry
Texas A&M University
tgentry@ag.tamu.edu

Adi Gevins
SoundVision Productions
adi.svslt@yahoo.com

David Gilbert
DOE Joint Genome Institute
degilbert@lbl.gov

Jeremy Glasner
University of Wisconsin
jglasner@wisc.edu

Tijana Glavina del Rio
DOE Joint Genome Institute
glavinadelrio1@llnl.gov

Daniela Goltsman
University of California, Berkeley
dgolts@eps.berkeley.edu

Elaine Gong
DOE Joint Genome Institute
elgong@lbl.gov

David Goodstein
DOE Joint Genome Institute
dmgoodstein@lbl.gov

Lynne Goodwin
Los Alamos National Laboratory
lynneg@lanl.gov

Stuart Gordon
Hiram College
gordonsg@hiram.edu

Igor Grigoriev
DOE Joint Genome Institute
ivgrigoriev@lbl.gov

Joe Grzymski
Desert Research Institute
joeg@dri.edu

Jenny Gu
University of Muenster
j.gu@uni-muenster.de

Liang Guo
Monsanto Company
liang.guo@monsanto.com

Christopher Hack
DOE Joint Genome Institute
cahack@lbl.gov

Matthew Hamilton
DOE Joint Genome Institute
mghamilton@lbl.gov

Nancy Hammon
DOE Joint Genome Institute
nmhammon@lbl.gov

Shunsheng Han
Los Alamos National Laboratory
han_cliff@lanl.gov

Miranda Harmon-Smith
DOE Joint Genome Institute
harmonsmith2@llnl.gov

Michael Harrington
UMBC
mharri2@umbc.edu

Amber Hartman
Johns Hopkins/UC Davis
alh@jhu.du

James Hartwell
University of Liverpool
hartwell@liv.ac.uk

Loren Hauser
ORNL
hauserlj@ornl.gov

Jennifer Hawkins
University of Georgia
jhawkins@uga.edu

David Hays
DOE Joint Genome Institute
dehays@lbl.gov

Ji He
The Samuel Roberts Noble Foundation
jhe@noble.org

Uffe Hellsten
DOE Joint Genome Institute
uhellsten@lbl.gov

Markus Herrgard
Synthetic Genomics, Inc.
markus.herrgard@gmail.com

Matthias Hess
DOE Joint Genome Institute
mhess@lbl.gov

David Hillman
DOE Joint Genome Institute
dwhillman@lbl.gov

Nathan Hillson
Stanford University
hillson@stanford.edu

Ann Hirsch
Univ of California, Los Angeles
ahirsch@ucla.edu

Emily Hollister
Texas A&M University
ehollister@tamu.edu

Kelli Hoover
Penn State University
kxh25@psu.edu

Attendees

Leila Hornick
DOE Joint Genome Institute
lahornick@lbl.gov

Ping Hu
Lawrence Berkeley National Lab
phu@lbl.gov

Donghui Huang
Exelixis Inc.
shuang@exelixis.com

Tom Huffaker
Piedmont Highlander
thuffaker@piedmont.k12.ca.us

Phil Hugenholtz
DOE Joint Genome Instittute
phugenholtz@lbl.gov

William Inskeep
Montana State University
binskeep@montana.edu

Lakshmi Jakkula
Lawrence Berkeley National Lab
lrjakkula@lbl.gov

Gabriel James
Australian National University
gabriel.james@anu.edu.au

Christer Jansson
Lawrence Berkeley National Lab
cgjansson@lbl.gov

Zack Jay
Montana State University
zackj@montana.edu

Cynthia Jeffries
Oak Ridge Nat'l Laboratory
jeffriescd@ornl.gov

Susan Jenkins
Energy Biosciences Insitute - UCB
sjenkins@berkeley.edu

Craig Johnson
BioVir Laboratories, Inc.
csj@biovir.com

Ginger Jui
University of California, Berkeley
ginger.jui@gmail.com

Marian Kaehler
Luther College
kaehlerm@luther.edu

Ulas Karaoz
Lawrence Berkeley National Lab
ukaraoz@lbl.gov

Steven Karpowicz
Univ of California, Los Angeles
skarp@chem.ucla.edu

David Keating
Great Lakes Bioenergy Research Center
dkeating@glbrc.wisc.edu

Lisa Kegg
DOE Joint Genome Institute
lrkegg@lbl.gov

Gert Kema
Plant Research International B.V.
gert.kema@wur.nl

Megan Kennedy
DOE Joint Genome Institute
mckennedy@lbl.gov

Richard Kerrigan
Sylvan Research
rwk@sylvaninc.com

Edward Kirton
DOE Joint Genome Institute
eskirton@lbl.gov

Valentin Klaus
AWI
valentin@awi.de

Dave Koppenaal
Pacific National Northwest Laboratory
david.koppenaal@pnl.gov

Frank Korzeniewski
DOE Joint Genome Institute
frkorzeniewski@lbl.gov

Anthony Kosky
DOE Joint Genome Institute
askosky@lbl.gov

Victor Kunin
DOE Joint Genome Institute
vkunin@lbl.gov

Cheryl Kuske
Los Alamos National Laboratory
kuske@lanl.gov

Sze Wai Kwok
The Chinese University of Hong Kong
kwokdorami@hotmail.com

Angie Lackey
Roche
angie.lackey@roche.com

Kris Lambert
University of Illinois
knlamber@uiuc.edu

Miriam Land
Oak Ridge National Laboratory
landml@ornl.gov

Robert Landick
UW Madison GLBRC
landick@bact.wisc.edu

Stephen Landt
Stanford University
sglandt@stanford.edu

Janna Lanza
Roche
janna.lanza@roche.com

Christina Lanzatella-Craig
USDA-ARS-WRRC
christina.craig@ars.usda.gov

Alla Lapidus
DOE Joint Genome Institute
alapidus@lbl.gov

Debbie Laudencia-Chingcuanc
USDA
debbie.laudencia@ars.usda.gov

Gaylynn LaVenture
DOE Joint Genome Institute
glaventure@lbl.gov

Janey Lee
DOE Joint Genome Institute
jlee2@lbl.gov

Patrick Lee
University of California, Berkeley
leep@berkeley.edu

Dawei Lin
UC Davis Genome Center
lhslin@ucdavis.edu

Tomas Linder
University of California, San Francisco
tlinder@cmp.ucsf.edu

Amelia Linnemann
Wayne State University - School of Medicine
amelia@compbio.med.wayne.edu

Mary Lipton
Pacific Northwest National Laboratory
mary.lipton@pnl.gov

Anna Lipzen
DOE Joint Genome Institute
alipzen@lbl.gov

Qiong Liu
Brookhaven National Lab
qliu@bnl.gov

Tiangang Liu
Stanford University
liutg@stanford.edu

Michelle Lizotte-Waniewski
University of Massachusetts
mlizotte@bio.umass.edu

Richard Long
University of South Carolina
rlong@biol.sc.edu

W. Lorenz
University of Georgia
wlorenz@uga.edu

Steve Lowry
DOE Joing Genome Institute
slowry@lbl.gov

Susan Lucas
DOE Joint Genome Institute
lucas11@llnl.gov

Athanasios Lykidis
DOE Joint Genome Institute
alykidis@lbl.gov

Eric Lyons
University of California, Berkeley
elyons@nature.berkeley.edu

Jon Magnuson
Pacific Northwest National Laboratory
jon.magnuson@pnl.gov

Stephanie Malfatti
DOE Joint Genome Institute
malfatti3@llnl.gov

Yuzuki Manabe
Joint BioEnergy Institute
ymanabe@lbl.gov

Vito Mangiardi
DOE Joint Genome Institute
vjmangiardi@lbl.gov

Victor Markowitz
Lawrence Berkeley National Lab
vmmarkowitz@lbl.gov

Konstantinos Mavrommatis
DOE Joint Genome Institute
kmavrommatis@lbl.gov

Patrick May
Max-Planck-Institute of Molecular Plant Physiology
may@mpimp-golm.mpg.de

Sarrah Ben M'Barek
Plant Research International
sarrah.benmbarek@wur.nl

Lee McCue
Pacific Northwest National Lab
leeann.mccue@pnl.gov

Tim McDermott
Thermal Biololgy Institute
timmcder@montana.edu

Sabeeha Merchant
Univ of California, Los Angeles
merchant@chem.ucla.edu

Joachim Messing
Rutgers University, Waksman Institute
messing@waksman.rutgers.edu

Folker Meyer
Argonne National Laboratory
folker@mcs.anl.gov

Natalia Mikhailova
DOE Joint Genome Institute
nmikhailova@lbl.gov

Kendra Mitchell
University of British Columbia
kendrami@interchange.ubc.ca

William Moe
Louisiana State University
moemwil@lsu.edu

Pia Moisander
University of California, Santa Cruz
pmoisander@pmc.ucsc.edu

Martin Mokrejs
Charles University
mmokrejs@iresite.org

Jenna Morgan
JGI/UC Davis
jlmorgan@lbl.gov

Jochen Mueller
Morgan State University
jochen.mueller@morgan.edu

Gerard Muyzer
Delft University of Technology
g.muijzer@tudelft.nl

Alexander Myburg
University of Pretoria
zander.myburg@fabi.up.ac.za

Campbell Nairn
University of Georgia
jnairn@warnell.uga.edu

Kemanthi Nandasena
Murdoch University
kemanthi@murdoch.edu.au

Simona Necula
DOE Joint Genome Institute
sfnecula@lbl.gov

Thong Nguyen
Applied Biosystems, Inc.
allen.nguyen@appliedbiosystems.com

Rita Nieu
USDA
rita.nieu@gmail.com

Krishna Niyogi
UC-Berkeley/LBNL
niyogi@nature.berkeley.edu

Henrik Nordberg
DOE Joint Genome Institute
hnordberg@lbl.gov

Donald Nuss
University of Maryland Biotechnology Institute
nuss@umbi.umd.edu

Take Ogawa
Roche Applied Science /454
take.ogawa@roche.com

John Ohlrogge
Michigan State Univ
ohlrogge@msu.edu

Miki Okada
U.S. Department of Agriculture, ARS
miki.okada@ars.usda.gov

Jeanine Olsen
University of Groniingen, Centre for Ecological & Evolutionary Studies
j.l.olsen@rug.nl

Jeffrey Osborne
Manchester College
jposborne@manchester.edu

Krishnaveni Palaniappan
BDMTC/IMG Team
kpalaniappan@lbl.gov

Jasmyn Pangilinan
DOE Joint Genome Institute
jlpangilinan@lbl.gov

Jacob Parnell
Utah State University
darlene.orduno@usu.edu

Matteo Pellegrini
Univ of California, Los Angeles
matteop@mcdb.ucla.edu

Ze Peng
DOE Joint Genome Institute
zpeng@lbl.gov

Christa Pennacchio
DOE Joint Genome Institute
cppennacchio@lbl.gov

Len Pennacchio
DOE Joint Genome Institute
lapennacchio@lbl.gov

Pamela Peralta-Yahya
LBL/JBEI
pperalta-yahya@lbl.gov

Rene Perrier
DOE Joint Genome Institute
raperrier@lbl.gov

Pavel Pevzner
Univ of California, San Diego
ppevzner@ucsd.edu

Wilson Phung
DOE Joint Genome Institute
wphung@lbl.gov

Anne Pinckard
California magazine
anne.pinckard@gmail.com

Attendees

Andre Pires da Silva
University of Texas at Arlington
apires@uta.edu

Samuel Pitluck
DOE Joint Genome Institute
s_pitluck@lbl.gov

Bradley Plantz
UNL
bplantz2@unl.edu

David Pletcher
Lawrence Berkeley National Laboratory
lpm_jgium@cathedralcanyon.net

Juergen Polle
Brooklyn College of CUNY
jpolle@brooklyn.cuny.edu

Amy Powell
Sandia National Laboratories
ajpowel@sandia.gov

Justin Powlowski
Concordia University
powlow@alcor.concordia.ca

Henry Priest
Oregon State University
priesth@onid.orst.edu

Simon Prochnik
DOE Joint Genome Institute
prochnik@gmail.com

Nan Qin
Beijing Genomics Institute
qinnan@genomics.org.cn

Preethi Ramaiya
NOVOZYMES INC
pira@novozymes.com

Beth Rasala
The Scripps Research Institute
brasala@scripps.edu

Wayne Reeve
Murdoch University
reeve@murdoch.edu.au

Jay Reichman
US EPANHEERL/WED
reichman.jay@epa.gov

Gary Resnick
Los Alamos National Lab
tanyal@lanl.gov

Sonia Reveco
DOE Joint Genome Institute
sareveco@lbl.gov

Kathryn Richmond
Great Lakes Bioenergy Research Center
kerichmond@wisc.edu

Frank Roberto
Idaho National Laboratory
francisco.roberto@inl.gov

Simon Roberts
DOE Joint Genome Institute
srroberts@lbl.gov

David Robinson
DOE Joint Genome Institute
dsrobinson@lbl.gov

Dan Rokhsar
DOE Joint Genome Institute
dsrokhsar@yahoo.com

Margie Romine
Pacific Northwest National Laboratory
margie.romine@pnl.gov

Giovanni Rompato
Utah State University
darlene.orduno@usu.edu

Pamela Ronald
UC Davis, Joint Bioenergy Institute
pcronald@ucdavis.edu

Eddy Rubin
DOE Joint Genome Institute
emrubin@lbl.gov

Doug Rusch
J. Craig Venter Institute
drusch@jcvi.org

Asaf Salamov
DOE Joint Genome Institute
aasalamov@lbl.gov

Christopher Sales
University of California, Berkeley
chris.sales@berkeley.edu

Wendy Sammons-Jackson
USAMRIID
wendy.sammons-jackson@amedd.army.mil

Laura Sandor
LBNL JGI
lcsandor@lbl.gov

Francisco Santoyo Santos
Public University of Navarre
patximotxo@yahoo.es

Wendy Schackwitz
DOE Joint Genome Institute
wsschackwitz@lbl.gov

Henrik Scheller
Lawrence Berkeley National Lab
hscheller@lbl.gov

Jeremy Schmutz
JGI-HudsonAlpha
jschmutz@hudsonalpha.org

Alexandra Schnoes
University of California, San Francisco
alexandra.schnoes@ucsf.edu

Ariel Schwartz
Synthetic Genomics
aschwartz@syntheticgenomics.com

Alexander Sczyrba
DOE Joint Genome Institute
asczyrba@lbl.gov

Margrethe Serres
Marine Biological Laboratory
mserres@mbl.edu

Nicole Shapiro
DOE Joint Genome Institute
nrshapiro@lbl.gov

Lucy Shapiro
Stanford University
shapiro@stanford.edu

Harris Shapiro
DOE Joint Genome Institute
hjshapiro@lbl.gov

Thomas Sharpton
Univeristy of California, Berkeley
sharpton@berkeley.edu

Christine Shewmaker
BluGoose Consulting
blugoose@sbcglobal.net

Steve Siembieda
Advanced Analytical
ssiembieda@aati-us.com

Shaneka Simmons
Jackson State University
shaneka.s.simmons@jsums.edu

Steven Singer
Lawrence Livermore National Lab
singer2@llnl.gov

Kanwar Singh
DOE Joint Genome Institute
ksingh@lbl.gov

Tatyana Smirnova
DOE Joint Genome Institute
tsmirnova@lbl.gov

Richard Smith
Pacific Northwest National Laboratory
richard.smith@pnl.gov

William Smith
Lawrence Livermore National Lab
smith324@llnl.gov

Chris Somerville
EBI
crs@berkeley.edu

Anton Sonnenberg
Wageningen UR
anton.sonnenberg@wur.nl

John Spear
Colorado School of Mines
jspear@mines.edu

Jason Stajich
University of California, Berkeley
jason_stajich@berkeley.edu

Michael Steinwand
USDA Agricultural Research Service
michael.steinwand@ars.usda.gov

Ioannis Stergiopoulos
Wageningen University
ioannis.stergiopoulos@wur.nl

Marvin Stodolsky
US Department of Energy
marvin.stodolsky@science.doe.gov

Garret Suen
University of Wisconsin-Madison
gsuen@wisc.edu

Hui Sun
DOE Joint Genome Institute
hsun@lbl.gov

Shinichi Sunagawa
University of California, Merced
ssunagawa@ucmerced.edu

Sirisha Sunkara
DOE Joint Genome Institute
ssunkara@lbl.gov

Bridget Swift
DOE Joint Genome Institute
bkswift@lbl.gov

Edward Szekeres
454 Sequencing/Roche
edward.szekeres@roche.com

Ernest Szeto
LBNL / JGI
eszeto@lbl.gov

Cristina Takacs-Vesbach
University of New Mexico
cvesbach@unm.edu

Yuhong Tang
The Samuel Roberts Noble Foundation
ytang@noble.org

Angela Tarver
DOE Joint Genome Institute
amtarver@lbl.gov

Caroline Taylor
Energy Biosciences Institute, UC Berkeley
cmtaylor@berkeley.edu

Gail Taylor
University of Southampton
g.taylor@soton.ac.uk

John Taylor
Univeristy of California, Berkeley
jtaylor@berkeley.edu

Shivegowda Thammannagowda
National Renewable Energy Laboratory
shivegowda_thammannagowda@nrel.gov

Damon Tighe
DOE Joint Genome Institute
tighe2@llnl.gov

Christian Tobias
USDA ARS Western Regional Research Center
christian.tobias@ars.usda.gov

Tamas Torok
Lawrence Berkeley National Lab
ttorok@lbl.gov

Susannah Tringe
DOE Joint Genome Institute
sgtringe@lbl.gov

Stephan Trong
DOE Joint Genome Institute
trong1@llnl.gov

Adrian Tsang
Concordia University
tsang@gene.concordia.ca

Stephen Turner
Pacific Biosciences
sturner@pacificbiosciences.com

Jerry Tuskan
ORNL/JGI
gtk@ornl.gov

Ludmila Tyler
University of California, Berkeley
ltyler@berkeley.edu

Jana Vaclavikova
Exelixis
jvaclavi@exelixis.com

Olivier Vallon
CNRS
ovallon@ibpc.fr

Daniel Van der Lelie
Brookhaven National Laboratory
vdlelied@bnl.gov

Kranthi Varala
Univ. of Illinois at Urbana-Champaign
kvarala2@uiuc.edu

Meric Velasco
DOE Joint Genome Institute
mvelasco@lbl.gov

J. Craig Venter
J. Craig Venter Institute
mtull@jcvi.org

Christopher Villalta
Taylor Lab - UC Berkeley
cvillalta@berkeley.edu

John Vogel
USDA ARS
john.vogel@ars.usda.gov

Henrik Von der Lippe
Lawrence Berkeleyl Natioanl Lab
hvdlippe@lbl.gov

Christian Voolstra
University of California, Merced
cvoolstra@ucmerced.edu

Valentina Vysotskaia
Exelixis
vvs@exelixis.com

Lawrence Wackett
University of Minnesota
wacke003@umn.edu

Setsuko Wakao
University of California, Berkeley
swakao@nature.berkeley.edu

Harkamal Walia
Univerisity of California, Davis
hwalia@ucdavis.edu

Vicki Walsworth
DOE Joint Genome Institute
vwalworth@lbl.gov

Hao Wang
The University of Georgia
wanghao@uga.edu

David Ward
Montana State University
umbdw@montana.edu

Falk Warnecke
DOE Joint Genome Insitute
fwarnecke@lbl.gov

Lidia Watrud
US EPA
watrud.lidia@epa.gov

Kimberlee West
Univeisity of California, Berkeley
kimberleew@gmail.com

Emily Whiston
UC Berkeley - PMB
whiston@berkeley.edu

Richard Whitaker
Enzymatics
whitaker@enzymatics.com

Curtis Wilkerson
Michigan State University
wilker13@msu.edu

Attendees

John Willis
Duke University
jwillis@duke.edu

Rod Wing
University of Arizona
rwing@ag.arizona.edu

Dana Wohlbach
University of Wisconsin-Madison
danawohlbach@gmail.com

Marcus Wohlsen
The Associated Press
mwohlsen@ap.org

Gordon Wolfe
California State Univ. Chico
gwolfe2@csuchico.edu

Man Chun Wong
The Chinese University of Hong Kong
manchun3@hotmail.com

Valerie Wong
University of California, Berkeley
vwong@berkeley.edu

Jason Wood
Montana State University
montanawoody@gmail.com

Tanja Woyke
DOE Joint Genome Institute
twoyke@lbl.gov

Crystal Wright
DOE Joint Genome Institute
cawright@lbl.gov

Cindy Wu
Lawrence Berkeley National Lab
chwu@lbl.gov

Dongying Wu
UC Davis Genome Center
dygwu@ucdavis.edu

Jiajie Wu
UADA ARS WRRC
jiajie.wu@ars.usda.gov

Denise Yadon
DOE Joint Genome Institute
dyadon@lbl.gov

Huihuang Yan
Great Lakes Bioenergy Research Center, University of Wisconsin
huihuangyan@wisc.edu

Rui Yang
Dynamic Throughput Inc.
tribur@gmail.com

Xiaohan Yang
Oak Ridge National Laboratory
yangx@ornl.gov

Suzan Yilmaz
DOE Joint Genome Institute
syilmaz@lbl.gov

Hugh Young
USDA, WRRC, GGD
hugh.young@ars.usda.gov

Mark Young
Montana State University
myoung@montana.edu

Heather Youngs
Energy Biosciences Institute
hlyoungs@berkeley.edu

Chunsheng Zhang
ArborGen LLC
cxzhang@arborgen.com

Jiyi Zhang
The Samuel Roberts Noble Foundation
jzhang@noble.org

Tao Zhang
DOE Joint Genome Institute
tzhang3@lbl.gov

Xueling Zhao
DOE Joint Genome Institute
xzhao@lbl.gov

Zhiying Zhao
DOE Joint Genome Institute
zyzhao@lbl.gov

Carol Zhou
Lawrence Livermore National Laboratory
zhou4@llnl.gov

Kemin Zhou
DOE Joint Genome Institute
kzhou@lbl.gov

Natasha Zvenigorodsky
DOE Joint Genome Institute
nzvenigorodsky@lbl.gov

Author Index

Authors

Authors

Notes