DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Office of Science
U.S. DEPARTMENT OF ENERGY

U.S. DEPARTMENT OF ENERGY

THIRD ANNUAL

U.S. Department of Energy
JOINT GENOME INSTITUTE

USER MEETING

MARCH 26–28, 2008
Marriott Hotel
Walnut Creek, California

An electronic version of this document is available at:
http://www.jgi.doe.gov/meetings/usermtg08/abstracts.html

All information current as of March 12, 2008

Meeting Coordinator
Marsha Fenner
DOE Joint Genome Institute
mwfenner@lbl.gov

The Joint Genome Institute is a user facility of
the U.S. Department of Energy Office of Science.

DOE Joint Genome Institute: ***www.jgi.doe.gov***
DOE Office of Science: ***science.doe.gov***

Third Annual
DOE Joint Genome Institute
User Meeting

Sponsored By

U.S. Department of Energy
Office of Science

March 26–28, 2008

Marriott Hotel

Walnut Creek, California

Contents

Agenda

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

WEDNESDAY, March 26

Start Time	*End Time*	*Subject*	*Session Chair/Speaker*
8:30 AM		***Registration Opens***	
9:00 AM	**12:00 PM**	**Workshops**	
		Eukaryotic Annotation Workshop **IMG Workshop** **JGI 101 Workshop**	
12:00 PM	1:00 PM	***Lunch provided for WORKSHOP PARTICIPANTS ONLY on this day***	
1:00 PM	**4:30 PM**	**Session 1- Biomass Feedstocks**	**Eddy Rubin, chair**
1:00 PM	1:10 PM	**Introduction**	**Eddy Rubin**
1:10 PM	1:40 PM	**JGI Plant Portfolio Overview**	**Jerry Tuskan**
1:40 PM	2:10 PM	**The *Sorghum bicolor* Genome, the Diversification of Cereals, and the Productivity of Tropical Grasses**	**Andrew Paterson**
2:10 PM	2:40 PM	**Miscanthus**	**Steve Long**
2:40 PM	3:00 PM	***Break***	
3:00 PM	3:30 PM	**Poplar Metabolism**	**Tim Tschaplinski**
3:30 PM	4:00 PM	***Eucalyptus*: Sequencing a Global Tree Genome for Energy, Fiber and Wood**	**Dario Grattapaglia**
4:00 PM	4:15 PM	***Short talk from poster abstracts***	Speaker TBD
4:15 PM	4:30 PM	***Short talk from poster abstracts***	Speaker TBD
5:30 PM	**6:30 PM**	**Welcome and Keynote Address** **The Helios Project**	**Steven Chu**
7:00 PM	10:00 PM	**Opening Reception and Poster Session**	

THURSDAY, March 27

Start Time	*End Time*	*Subject*	*Session Chair/Speaker*
8:30 AM	**10:00 AM**	**Session II – Plant cell wall**	**Dan Rokhsar, Chair**
8:30 AM	9:00 AM	**Plant Cell Walls: The Biomass for Ethanol Production**	**Debra Mohnen**
9:00 AM	9:30 AM	**Genomics Approaches to Maize Anther Development**	**Virginia Walbot**
9:30 AM	10:00 AM	***Break***	

10:00 AM	**11:30 AM**	**Session III – Microbes for Bioenergy**	**Phil Hugenholtz, Chair**
10:00 AM	10:30 AM	**Biomass Recalcitrance: Barrier to Economic Ethanol Biorefineries**	**Mike Himmel**
10:30 AM	11:00 AM	**The Search for Microbial Consortia for Cellulose Conversion**	**Martin Keller**
11:00 AM	11:30 AM	**Biofuels Synthesis**	**James Liao**
11:30 AM	**1:00 PM**	***Lunch***	
12:00 PM	**1:00 PM**	**Business Meeting**	
1:30 PM	**5:00 PM**	**Session IV- Microbial Genomics**	**Nikos Kyrpides, Chair**
1:30 PM	2:00 PM	**Systems Approach to Microbial Evolution**	**Bernhard Palsson**
2:00 PM	2:30 PM	**Life in the Slow Lane: Ecogenomics of an Extreme Environment**	**Terry Hazen**
2:30 PM	3:00 PM	**Comparative Fungal Genomics: *Batrachochytrium* and *Neurospora***	**John Taylor**
3:00 PM	3:30 PM	***Break***	
3:30 PM	4:00 PM	**Microbial Diversity and the "Rare Biosphere"**	**Mitch Sogin**
4:00 PM	4:30 PM	**Genomic Approaches to Engineering Fungal Metabolism and Stress Tolerance**	**Audrey Gasch**
4:30 PM	4:45 PM	***Short talk from poster abstracts***	Speaker TBA
4:45 PM	5:00 PM	***Short talk from poster abstracts***	Speaker TBA
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 28

Start Time	*End Time*	*Subject*	*Session Chair/Speaker*
9:00 AM	**12:00 PM**	**Session V- Emerging Technologies**	**Jim Bristow, Chair**
9:00 AM	9:30 AM	Algae for Biofuels	**Steve Mayfield**
9:30 AM	10:00 AM	**Bacterial Genome Synthesis and Transplantation: Progress on Constructing a Synthetic Cell**	**John Glass**
10:00 AM	10:30 AM	***Break***	
10:30 AM	11:00 AM	**Massively Parallel Resequencing**	**Jay Shendure**
11:00 AM	11:30 AM	**Analysis of the Significance of Fine-scale Diversity Within Natural Microbial Populations**	**Jill Banfield**
11:30 AM	12:00 PM	**Short Read Technologies at JGI**	**Len Pennacchio**
12:00 PM		***End of User Meeting***	

Speaker Presentations

Abstracts alphabetical by speaker

Acid Mine Community Genomics

Jill Banfield (jbanfield@berkeley.edu)

University of California, Berkeley, California

Genomic Approaches to Engineering Fungal Metabolism and Stress Tolerance

Audrey Gasch (agasch@wisc.edu)

University of Wisconsin, Madison, Wisconsin

S. cerevisiae is the current workhorse in industrial ethanol production, however maximal conversion of biomass to biofuel is hindered due to a number of limitations. The response of yeast to stressful fermentation conditions, including elevated temperature, low pH, high osmolarity, and toxic ethanol concentrations, caps the amount of alcohol that strains can produce and limits the effectiveness of simultaneous saccharification and fermentation. The utility of *S. cerevisiae* in fermenting alternate feedstock is also limited since this species cannot inherently ferment pentose sugars, a major component of hemicellulosic material. The Great Lakes Bioenergy Research Center (GLBRC) seeks to engineer multi-stress resistant, xylose-fermenting 'super yeast' to overcome these limitations. A key feature of our approach is to exploit natural variation and evolution to identify mechanisms of stress defense and xylose metabolism. Functional and comparative genomic approaches will be discussed.

Bacterial Genome Synthesis and Transplantation: Progress on Constructing a Synthetic Cell

John I. Glass (jglass@jcvi.org), Carole Lartigue, Daniel G. Gibson, Gwynedd A. Benders, Clyde A. Hutchison III, Hamilton O. Smith, and J. Craig Venter

The J. Craig Venter Institute, Rockville, Maryland

Mycoplasma genitalium is an approximately 300nm diameter wall-less bacterium that has the smallest known genome of any cell that can be grown in pure culture.. When this human urogenital pathogen is grown in the laboratory in a rich, serum-containing medium, as at least 110 or its 485 protein coding genes are not essential based on one-gene-at-a-time transposon mutagenesis. In order to better understand the essence of a minimal cell, we employing a synthetic genomics approach to construct a 582,970 bp *M. genitalium* genome. The synthetic genome will contain all the genes of wild type *M. genitalium* G37 except MG408, which will be disrupted by an antibiotic resistance marker to block pathogenicity and to allow for selection. Overlapping "cassettes" of 5-7 kb, assembled from chemically synthesized oligonucleotides, are being joined by *in vitro* recombination to produce intermediate assemblies of approximately 24 kb, 72 kb ("1/8 genome"), and

144 kb ("1/4 genome") and cloned as bacterial artificial chromosomes (BACs) in *Escherichia coli*. Once assemblies of all four 1/4 genomes are identified, the complete synthetic genome will be assembled and cloned in the yeast *Saccharomyces cerevisiae*. Minimization of the synthetic genome can be carried out by assembly of cassettes with individual genes deleted or by genome reduction using recombineering methods. Both approaches require the development of methods to transplant the synthesized genome into a receptive cytoplasm, such that the donor genome becomes installed as the new operating system of the cell. As a step toward propagation of synthetic genomes, we completely replaced the genome of *Mycoplasma capricolum* with one from *Mycoplasma mycoides* LC by transplanting a whole genome as naked DNA. These cells that result from genome transplantation are phenotypically identical to the *M. mycoides* LC donor strain as judged by several criteria. We believe these synthetic genomics methods will lead eventually to design and construction of new cells endowed with capacities to address human needs in medicine, energy, and environmental remediation.

Eucalyptus: Sequencing a Global Tree Genome for Energy, Fiber and Wood

Alexander A. Myburg,[1] **Dario Grattapaglia**,[2,3] Gerald A. Tuskan,[4] Jeremy Schmutz,[5] Daniel S. Rokhsar,[6,7] Kerrie Barry,[7] Jim Bristow,[6] and The *Eucalyptus* Genome Network (EUCAGEN)[8]

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa; [2]EMBRAPA Genetic Resources and Biotechnology, Estação Parque Biológico, Brasilia, Brazil; [3]Graduate Program in Genomic Sciences and Biotechnology, Universidade Catolica de Brasília-SGAN, Brasilia, Brazil; [4]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Stanford Human Genome Center, Stanford University, Palo Alto, California; [6]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California; [7]DOE Joint Genome Institute, Walnut Creek, California; and [8]www.eucagen.org

A key step for the achievement of a sustainable energy future is our understanding of the molecular basis of superior growth and adaptation in woody plants suitable for biomass production. A truly "global" tree, the genus *Eucalyptus* includes over 700 different species native to Australia and islands to its north where they occur naturally from sea level to the alpine tree line, from high rainfall to semiarid zones, spanning a wide environmental gradient from latitude 3° to 43° south. Due to their wide adaptability and excellent wood properties, eucalypts were rapidly adopted for plantation forestry around the world since their discovery in the 18th century. Today, *Eucalyptus* tree species are among the most widely planted and fastest growing woody plants in the world. Almost 20 million hectares are planted in 90 countries reaching mean annual growth increments four times those of conifers. The biomass production and carbon sequestration capacities of *Eucalyptus* trees are aligned with the DOE missions of alternative energy production and global carbon cycling. Eucalypts sequester carbon at an average rate of 10 tons of carbon/ha/yr from planting to harvesting and up to 14 tons/ha/yr in fast-growing tropical plantations. Furthermore, eucalypts have a positive net carbon balance even after computing the production of CO_2 when used for energy from charcoal or as pulp and paper. Their ability to produce up to to 100 m^3/ha/year of cellulose-rich wood fiber makes fast-growing eucalypt species premier candidates for the renewable production of lignocellulose biomass for ethanol production. The U.S. DOE Joint Genome Institute will sequence the genome of *Eucalyptus grandis* a candidate biomass energy crop. This is the result of a

successful proposal by the *Eucalyptus* Genome Network (EUCAGEN, www.eucagen.org), an international network of more than 130 scientists in 18 countries. In parallel to a whole genome shotgun, BAC end and ESTs sequencing from the target genome done at JGI and JGI-SHGC, EUCAGEN members are contributing with a number of genomic resources such as 30X coverage BAC libraries, high-density linkage maps, a multi-species EST database with more than one million sequences, microarray resources and genetic transformation systems. The estimated genome size of *E. grandis* is ~640 Mbp. It is a diploid species with a haploid chromosome number of n = 11. Due to its preferentially outcrossing mating system and nascent domestication, *E. grandis* displays a high level of nucleotide diversity (~ 1%) and frequent indels throughout the genome. To mitigate the expected challenges for assembling such a genome, a *E. grandis* tree (BRASUZ1) derived from one generation of selfing was chosen as the target and efforts are underway to develop and eventually sequence haploid tissue lines to aid haplotype discrimination. Only the second forest tree to be sequenced, *Eucalyptus* offers extraordinary opportunities for comparative genomic analysis with *Populus*, *Vitis* and with herbaceous species such as *Arabidopsis* and rice. All these attributes together with its unique evolutionary history, keystone ecological status and adaptation to marginal sites makes *Eucalyptus* an excellent focus for expanding our knowledge of the evolution and adaptive biology of woody perennials.

Life in the Slow Lane: Ecogenomics of an Extreme Environment

Terry C. Hazen[1,2] (TCHazen@lbl.gov) and Dylan Chivian[1,2]

[1]Virtual Institute for Microbial Stress and Survival (http://vimss.lbl.gov), and [2]Lawrence Berkeley National Laboratory, Berkeley, California

A more complete picture of life on Earth, and even life *in* the Earth, has recently become possible through the application of environmental genomics. We have obtained the complete genome sequence of a new genus of the *Firmicutes*, the uncultivated sulfate reducing bacterium *Desulforudis audaxviator*, by filtering fracture water from a borehole at 2.8 km depth in a South African gold mine. The DNA was sequenced using a combination of Sanger sequencing and 454 pyrosequencing, and assembled into just one genome, indicating the planktonic community is extremely low in diversity. We analyzed the genome of *D. audaxviator* using the MicrobesOnline annotation pipeline and toolkit (http://www.microbesonline.org), which offers powerful resources for comparative genome analysis, including operon predictions and tree-based comparative genome browsing. MicrobesOnline allowed us to compare the *D. audaxviator* genome with other sequenced members of the *Firmicutes* in the same clade (primarily *Pelotomaculum thermoproprionicum*, *Desulfotomaculum reducens*, *Carboxydothermus hydrogenoformans*, and *Moorella thermoacetica*), as well as other known sulfate reducers and thermophilic organisms. *D. audaxviator* gives a view to the set of tools necessary for what appears to be a self-contained, independent lifestyle deep in the Earth's crust. The genome is not very streamlined, and indicates a motile, endospore forming sulfate reducer with pili that can fix its own nitrogen and carbon. *D. audaxviator* is an obligate anaerobe, and lacks obvious homologs of many of the traditional O_2 tolerance genes, consistent with the low concentration of O_2 in the fracture water and its long-term isolation from the surface. *D. audaxviator* provides a complete genome representative of the Gram-positive bacteria to further our understanding of dissimilatory sulfate reducing bacteria and archaea. Additionally, study of the deep subsurface has offered access to the simplest community yet studied by environmental genomics, perhaps consisting of just a single species that is capable of performing all of the tasks necessary for life.

Biomass Recalcitrance: Barrier to Economic Ethanol Biorefineries

Michael Himmel (Mike_Himmel@nrel.gov), William Adney, Shi-You Ding, David Johnson, Michael Crowley, Mark Nimlos, and Thomas Foust

National Renewable Energy Laboratory, Golden, Colorado

Lignocellulosic biomass has long been recognized as a potential low-cost source of mixed sugars for fermentation to fuel ethanol. Several technologies have been developed over the past 80 years that allow this conversion process to occur, often in wartime context, yet the clear objective now is to make this process cost competitive in today's markets. Replacing 30% of U.S. 2004 finished motor gasoline demand (or about 60 billion gallons) with ethanol by 2030 will require a significant increase in ethanol production over today's corn starch-based industry. This process is technically feasible for corn stover and wheat straw today using biochemical conversion technology that includes pretreatment, enzymatic hydrolysis, and fermentation. However, the process remains fundamentally inefficient and is therefore risky to commercialize. Indeed, in order to ensure a successful transition from existing to 2030 technologies, investing in knowledge-based solutions to critical barriers is essential.

An important near term research strategy is to pursue deep understanding of natural plant cell-wall biosynthesis and degradation processes. Plant cell walls are composed mainly of lignocellulose, complex polymers constructed from simple sugars and aromatic monomers. Researchers now propose to explore the biological, biochemical, and structural properties of the wall synthesis and hydrolytic enzymes identified from the genomes of fungi, microbes, and plants; as well as the meta-genomes of plant matter decay communities.

Feedstock costs are now a major component of the ethanol commodity product price. Therefore, yield of lignocellulose-derived sugars is perhaps of highest priority. Yield issues touch many other critical biorefinery operations. Another impact on feedstock yield is associated with cellulases and other polysaccharide-degrading enzymes. Currently, very high loadings of cellulases are needed to reach 95% conversion of cellulose in pretreated biomass after 3 to 5 days using the simultaneous saccharification & fermentation (SSF) process. It is clear that only through an understanding of cellulase action at the molecular scale can improvements be made which reduce the enzyme cost.

Trichoderma reesei cellobiohydrolase I is a key cellulase enzyme in current commercial bioenergy enzyme preparations. This enzyme is truly a protein machine, as it is thought to "process" from the reducing end of a cellodextrin chain in crystalline cellulose, producing glucose and cellobiose as it moves. To assist in understanding the complexity of this unique enzymatic mechanism, we have employed molecular dynamics simulations using CHARMM and LAMMPS. These simulations include a model of cellulose 1β and the *T. reesei* cellobiohydrolase I enzyme interacting with the 1,0,0 cellulose surface, all enclosed in a box of water molecules. Initial simulations support a surprising new mechanism for the cellulose binding domain, as well as new hints about the overall functionality of the catalytic domain.

The Search for Microbial Consortia for Cellulose Conversion

Martin Keller (kellerm@ornl.gov)

BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee

The challenge of converting cellulosic biomass to sugars is the dominant obstacle to cost-effective production of biofuels in sustained quantities capable of impacting U.S. consumption of fossil transportation fuels. The BioEnergy Science Center (BESC) research program will address this challenge with an unprecedented interdisciplinary effort focused on overcoming the recalcitrance of biomass. By combining engineered plant cell walls to reduce recalcitrance with new biocatalysts to improve deconstruction, BESC within five years will revolutionize the processing of biomass.

The development of consolidated bioprocessing, which involves the use of a single microorganism or microbial consortium to overcome biomass recalcitrance through single-step conversion of biomass to biofuels may require the discovery and isolation of novel microorganisms and their genes. The extension of Norm Pace's cultivation-independent molecular phylogenetic approach to shotgun sequence whole environments followed by an increasing repertoire of tools for "post-environmental genomics" has lead to the realization, that the microbial diversity found in almost all investigated environments is much larger than ever anticipated. This tremendous diversity in combination with the finding of significant lateral gene transfer within these environments challenges the conventional understanding and definition of a microbial species and their function within the environment. In nature, cellulose utilization is carried out not by pure cultures of microorganisms but by multiple cellulolytic species coexisting with each other and with many noncellulolytic species. Most microorganisms thought to play an important role in cellulose hydrolysis in nature have evolved strategies that bring the cell close to the cellulose surface and give the cellulolytic organism first access to hydrolysis products leading to the formation of potentially complex microbial biofilms. The general mechanisms of adhesion at the point of contact to cellulose have been identified for several anaerobic cellulolytic species which includes the cellulososme organelle, noncatalytic cellulose-binding proteins, glycosylates moieties and pilus-like structures. The characterization of complex microbial environments will require new cultivation methods and detection tools specialized for the enrichment and analysis of these types of microbial consortia, supplementing metagenomic analyses. This presentation will outline how BESC will utilize metagenomics in conjunction with novel cultivation and analytical tools to exploit nature's diversity.

Next Generation Biofuels

James Liao (liaoj@ucla.edu)

University of California, Los Angeles, California

Miscanthus

Steve Long (slong@uiuc.edu), Matthew Hudson, Ray Ming, Steve Moose, and Tom Voigt

Energy Biosciences Institute, University of Illinois, Urbana, Illinois

Miscanthus is a genera of C4 grasses, closely related to both *Saccharum* and *Sorghum*, but including species which are exceptionally cold tolerant. *Miscanthus x giganteus*, a triploid hybrid of *Miscanthus sinensis* and *M. sacchariflorus* has proved highly productive in much of western Europe, and now in Illinois. Yields appear double those of switchgrass cultivars. This presentation will review why this biofuel feedstock is so productive and resource use efficient and the opportunities for improvement. Although the genetics of *Miscanthus* is poorly understood at present its close relationship with *Sorghum* and *Saccharum* should accelerate development of genomic resources. The strategies that are being used in the Energy Biosciences Institute to develop these will be outlined.

Algae for Biofuels

Stephen P. Mayfield (mayfield@scripps.edu)

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

Algae can produce biomass at more than ten times the rate of terrestrial plants on a unit area basis, making algae a potentially significant and economically viable source of sustainable bio-energy. Due to issues related to arable land, fresh water, nutrient loading of waterways, and N_2O loading of the atmosphere, bio-energy from row crops face significant challenges to provide energy at the scales required without sever consequences on food production. Algae are extremely efficient at nutrient uptake and can be grown in waste or marginal water on non-arable land, and can also be used to efficiently sequester carbon dioxide released from fossil fuel power plants. The attributes make algae a viable alternative for biofuel production in this country in a sustainable manner that does not detract from food production. Algae have been grown commercially at large scale, and naturally produce energy dense molecules like fatty acids and hydrocarbons. In order to fully exploit the biological advantages of algae we need to develop the knowledge base, as well as the genetic and molecular tools, and industrial processes necessary to enable algae oil production at a meaningful scale. Key to developing the tools and knowledge base required to achieve these goal will be the genomic, proteomic and metabolomic characterization of model algal species. The choice of algal species will depend upon a number of factors including growth rates, desired products, and biochemical and genetic characteristics. Characteristics of existing algal model species will be presented as well as a discussion of the attributes that need to be considered in choosing additional algal species for investigation as biofuel platforms.

Plant Cell Walls: The Biomass for Ethanol Production

Debra Mohnen (dmohnen@ccrc.uga.edu)

University of Georgia, Athens, Georgia

The transition from fossil fuels to biofuels takes advantage of the ability of plants to convert carbon dioxide and water in the presence of sunlight into diverse hydrocarbons. The emergent use of plant lignocellulosic biomass (i.e. plant cell walls) to produce bioethanol relies on the facile and economic conversion of plant wall polysaccharides into monosaccharides or small oligosaccharides that can be fermented into ethanol. Wall polysaccharides are a diverse and structurally complex group of polymers that serve a broad range of developmental, structural and defense functions in plants. The wall polysaccharides consist of cellulose, hemicelluloses and pectins that are embedded in a partially cross-linked polysaccharide matrix containing proteins, and in many secondary walls, lignin. A current challenge to economic bioethanol production from plant biomass is the difficulty of extracting the full monosaccharide potential from the wall matrix, i.e. the recalcitrance of walls to deconstruction. One strategy to reduce biomass recalcitrance is to modify plant wall synthesis to yield plants with structurally functional walls that are more easily deconstructed. Efforts to achieve this goal make clear the need for a deeper understanding of the molecular mechanisms that control plant cell wall synthesis and plant structure. The genes involved in the production, remodeling and regulation of plant cell wall synthesis include those encoding transcription factors, substrate biosynthetic enzymes, polysaccharide biosynthetic and remodeling enzymes, as well as proteins involved in wall cross-linking and lignin synthesis. In this talk the structure of plant primary and secondary walls will be presented and the current level of understanding of wall biosynthesis and wall biosynthetic genes will be summarized.

Systems Approach to Microbial Evolution

Bernhard Palsson (bpalsson@ucsd.edu)

University of California, San Diego, California

The *Sorghum bicolor* Genome, The Diversification of Cereals, and The Productivity of Tropical Grasses

Andrew H. Paterson (paterson@uga.edu)

University of Georgia, Athens, Georgia

Tropical grasses are among the most efficient biomass accumulators known, thanks to 'C4' photosynthesis, a complex combination of biochemical and morphological specializations discovered in sugarcane that confer efficient carbon assimilation at high temperatures. The Saccharinae clade of tropical grasses is of singularly-large importance, including three leading candidate lignocellulosic biofuels crops, *Sorghum* (currently the #2 U.S. biofuels crop), *Saccharum* (sugarcane and its relatives, currently the #1 biofuels crop worldwide), and *Miscanthus*. Its small genome (~730 Mb) and low level of gene duplication comparable to rice makes *Sorghum* an attractive model for functional genomics of C4 grasses in general and the Saccharinae in particular, and motivated its complete sequencing by the U.S. Department of Energy Joint Genome Institute (JGI) 'Community Sequencing

Program'. Using a whole-genome-shotgun sequencing approach together with paired-end backfilling and reconciliation with genetic and physical maps, the 201 largest scaffolds cover 97.3% of nucleotides, with 89.7% anchored to chromosomes. Comparison of sorghum to rice and other cereals illuminates the nature of their common ancestor, and the genes and mechanisms that have contributed to their structural and functional diversification. Further detailed study of the Saccharinae clade offers benefits ranging from improvement of its biofuels crops, to better understanding of its weedy/invasive members.

Massively Parallel Resequencing

Jay Shendure (shendure@u.washington.edu)

University of Washington, Seattle, Washington

Second-generation sequencing technologies have reduced the cost of DNA sequencing by over two orders of magnitude. Although assembled reference genomes for many species of interest are now available, resequencing, i.e. sequencing an individual(s) with the goal of identifying genetic variation, remains an important goal. It is frequently the case that investigators are interested in identifying germline variation or somatic mutations in a particular subset of a genome. However, our ability to take advantage of the power of new sequencing technologies is markedly impaired by the lack of a corresponding targeting method, analogous to PCR, that is matched to the scale at which the new sequencing platforms routinely operate. To meet this need, we are exploring several strategies for multiplex capture of complex genomic subsets. I will discuss the use of such "genome partitioning" methods in combination with second-generation sequencing to perform targeted variation discovery.

Microbial Diversity and the "Rare Biosphere"

Mitchell L. Sogin (sogin@mbl.edu)

Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts

Sequence analyses of nearly full-length ribosomal RNAs have revealed that microbial diversity is at least 100-1000 times greater than estimates based upon cultivation-dependent surveys. Yet these molecular assays have only captured a fraction of microbial diversity and they rarely provide estimates of relative abundance for different kinds of microbes or operational taxonomic units (OTUs). The promise of discovering new phylotypes has fostered experimental designs where low resolution procedures e.g. restriction fragment length polymorphisms - identify putatively distinct clones for DNA sequencing. This binning procedure comes at the expense of minimizing information about the relative abundance of distinct phylotypes. Typical molecular surveys that collect 1000-10,000 sequences describe a small fraction of the 10^8 microbes/liter commonly found in aquatic environments (fewer than one per million cells). Dominant populations have masked the detection of low-abundance organisms and some of these may be the most interesting for predicting shifts in microbial population structures in response to ecological change. The recent introduction of massively parallel pyrosequencing technology makes possible the collection of 300,000-450,000 short "tag" sequences from rRNA hypervariable regions in a single pyrosequencing run. Each serves as a proxy for the occurrence of a microbe in a community. The tag data sets provide first-order descriptions of the kinds and relative abundances of distinct OTUs. With this technology we have targeted and compared hypervariable regions V3 and V6. The relative abundance of

different OTUs in these data sets varied by more than three orders of magnitude. A relatively small number of tag sequences represent dominant bacterial populations. The vast majority (75%) of the tag sequences are very similar to each other and to entries in the reference database. Underlying the major populations is a broad distribution of distinct bacterial taxa that represent extraordinary diversity. These highly divergent, low-abundance organisms constitute a "rare biosphere" that is largely unexplored. Some of its members might serve as keystone species within complex consortia, others might simply be the products of historical ecological change with the potential to become dominant in response to shifts in environmental conditions that favor their growth. Because we know so little about the global distribution of members of the rare biosphere it is unknown if they represent specific biogeographical distributions of bacterial taxa, functional selection by particular marine environments, or cosmopolitan distribution of all microbial taxa – the "everything is everywhere" hypothesis.

Comparative Fungal Genomics: *Batrachochytrium* and *Neurospora*

Jason E. Stajich, Thomas J. Sharpton, Christopher Ellison, David Jacobson, N. Louise Glass, Donald O. Natvig, Erica Bree Rosenblum, Michael B. Eisen, and **John W. Taylor** (jtaylor@nature.berkeley.edu)

University of California, Berkeley, California

Sequencing of the first fungus from a very basal fungal clade, *Batrachochytrium dendrobatidis*, and of two fungi closely related to the model fungus *Neurospora crassa*, *N. tetrasperma* and *N. discreta*, highlights the value of comparative fungal genomics. Comparison of *B. dendrobatidis* to other fungi identified genes associated with previously known differences in motility and nuclear division, but also indicated that *B. dendrobatidis* cell wall composition must be different from that of more recently derived fungi. Use of cell wall chemistry to challenge this hypothesis has left intact the computational prediction. Knowledge that this fungus lives in animals, whereas its relatives are associated with plants, led to hypotheses about gene families whose expansion would be consistent with a diet of animal proteins, such as, keratin. Computational discovery of gene family expansion failed to falsify this hypothesis. Sequencing of two additional *Neurospora* species immediately brought the power of comparative genomics to problems of assembly and annotation that remained for one of the first and best studied fungal genomes, *Neurospora crassa*. With improved assemblies and annotation for the three *Neurospora* genomes, we now are able to investigate gene family expansion and contraction, synteny and rearrangement, and the effects of selection.

Poplar Metabolism

Timothy J. Tschaplinski (tschaplinstj@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

JGI Plant Portfolio Overview

Jerry Tuskan (tuskanga@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

Genomics Approaches to Maize Anther Development

Virginia Walbot (WALBOT@stanford.edu)

Stanford University, Palo Alto, California

Maize (*Zea mays L.*) is the model C4 genetic species and is also a very close relative of C4 grasses that are current leading candidates for biofuel production. Translational genomics from maize to biofuel crops should provide a "rapid start" to genomics approaches in these non-model plants. As an example of how maize biology can provide the context for focused studies in non-model relatives, the case study of maize anther development will be presented. The stamen is the "male" organ of the angiosperm flower. This compound organ has two parts, a narrow stem-like filament and a terminal anther. Maize anthers have four compartments (locules) each with just 5 cell types; the central most cells undergo meiosis and the resulting haploid cells develop into pollen. Our focus is understanding the genetic programming and cellular events that allow cell fate acquisition and cell fate maintenance within the early developmental stages of the anther up to successful entry into meiosis. Using mutants that are disrupted in anther cell fate acquisition and both transcriptome and proteome profiling we are beginning to understand how anther development is controlled. One of the most striking findings is that over the course of approximately 1 week, anthers express ~35,000 genes, about 70% of the expected gene number in maize. The relatively simple anther expresses nearly twice as many genes as leaves.

Poster Presentations

Posters alphabetical by first author. *Presenting author.

Adaptations to Iron Stress in a Pennate Marine Diatom: Global Cellular Response and Comparative Genomics

Andrew E. Allen[1,5]* (aallen@jcvi.org), Uma Maheswari,[1] Markus Lommer,[2] Alisdair Fernie,[3] Chis Bowler,[1,4] and Julie LaRoche[2]

[1]UMR8186, Dept of Biology, Ecole Normale Supérieure, Paris, France; [2]Institut fuer Meereskunde, Germany; [3]Max Planck Institute of Molecular Plant Physiology, Golm, Germany; [4]Cell Signaling Laboratory, Stazione Zoologica, Villa Comunale, Naples, Italy; and [5]J. Craig Venter Institute, Rockville, Maryland

It has become increasingly clear that iron (Fe) availability plays a major role in regulating the fate of upwelled nitrate (NO_3^-) and determining the size structure and community composition of phytoplankton assemblages in open-ocean and coastal upwelling regions. All of the Fe enrichment experiments conducted to date have reported sharp increases in the biomass and photosynthetic capacity of diatoms. Mounting evidence from field experiments, detailed physiological investigation, and genomic sequence data suggest fundamental differences in Fe bioavailability and uptake mechanisms, storage capacity, and stress recovery between pennate and centric diatoms. Pennate diatoms often dominate the phytoplankton assemblage after mesoscale Fe addition experiments because, in part, they are able to maintain cell viability during long periods of chronic Fe stress. The underlying molecular bases for these adaptations are virtually unknown. Large scale analyses of gene expression and gas chromatograph-mass spectroscopy (GC-MS) primary metabolite profiling data of Fe-limited *Phaeodactylum tricornutum* suggest that major metabolic reconfigurations are necessary to meet increased demand for Fe-stress metabolites such as those involved in reactive oxygen species (ROS) defense and intracellular metal chelation. Cellular nitrogen (N) status, and the accumulation of glutamate in particular, appear likely to play a primary role in remodeling of the photosynthetic apparatus in order to facilitate rapid recovery from Fe stress. Other major cellular adjustments to Fe stress include an overall down-regulation of photosynthesis, N and C reallocation from protein and storage carbohydrate degradation, and removal of excess electrons by mitochondrial alternative oxidase (AOX). The availability of genome data for centric and pennate diatoms provides a basis for beginning to understand the mechanisms by which pennate diatoms dominate over centric diatoms in chronically Fe starved high nutrient low chlorophyll (HNLC) waters.

Metagenomic Sequencing of Geothermal Microbial Communities Provides Insights Regarding Phylogenetic and Functional Diversity Within and Across Extreme Environments

J. Badger,[1] M. Bateson,[2] E. Boyd,[2] B. Fouke,[3] M. Frazier,[1] G. Geesey,[2] D. Haft,[1] N. Hamamura,[4] **W. Inskeep**[2]* (binskeep@montana.edu), Z. Jay,[2] M. Kozubal,[2] R. Macur,[2] A. Ortmann,[2] A-L. Reysenbach,[4] F. Roberto,[5] D. Rusch,[1] and M. Young[2]

[1]J. Craig Venter Institute, Rockville, Maryland; [2]Thermal Biology Institute, Montana State University, Bozeman, Montana; [3]Institute for Genomic Biology, University of Illinois,

Urbana, Illinois; [4]Department of Biology, Portland State University, Portland, Oregon; and [5]Biological Systems Department, Idaho National Laboratory, Idaho Falls, Idaho

Metagenomic sequence analysis of microbial communities holds promise for understanding prokaryotic gene content and diversity as a function of environmental parameters in natural, complex systems. Moreover, as sequencing costs decline, direct sequencing of environmental DNA may prove to be an efficient method for understanding the genetics of an enormous number of uncultured microorganisms. Ultimately the success of metagenomic approaches will hinge in part on the trade-off between microbial diversity (i.e. genomic diversity) and the amount of sequencing required to achieve adequate coverage and representation of dominant phylogenetic and functional attributes. High-temperature geothermal environments of Yellowstone National Park (YNP) generally exhibit strong selective forces, including extreme pH and or the predominance of specific electron donors and acceptors. As a result, these ecosystems offer a tremendous opportunity for applying metagenomic sequence analyses to microbial communities that are often considered less diverse, and often described as *model* communities. The goals of this project were to evaluate the phylogenetic and functional gene diversity of five geochemically diverse, high-temperature chemotrophic (i.e. non-photosynthetic) communities in YNP using modest metagenomic sequencing and bioinformatic analyses. Prior work on each of the sites including 16S rRNA gene sequence analysis and detailed geochemical characterization provided environmental context useful in site selection and in data interpretation. Based on our prior estimates of the number of dominant phylotypes in these communities, we hypothesized that 10-15 Mbases of random shotgun sequencing would be sufficient to establish a significant fraction of the phylogenetic and functional diversity represented in these thermophilic habitats, while understanding that this would not be sufficient coverage for assembling consensus genomes. Consequently, for each of the five sites included in this study, approximately 14,000 random sequence reads (averaging ~650 nt) were obtained from small-insert clone libraries prepared from environmental DNA. Phylogenetic analysis of individual sequence reads across sites demonstrated that (i) each site is dominated by a limited number of phylogenetic groups, and (ii) the metagenomic sequence data are consistent with prior 16S rRNA gene analyses and geochemical context for each site. The geochemical context is critical for understanding the phylogenetic singularity of specific sites, as well as the phylogenetic similarity occurring between specific sites. A significant number of individual sequence reads were assembled into larger scaffolds for all sites, ranging from as high as 80% of the sequence reads for an Aquificales dominated site to as low as ~30% of the reads for an acidic Fe-oxyhydroxide microbial mat. Functional analyses of the metagenomic sequence content also yielded important signatures that we hypothesize will be consistent with *system defining* physiochemical processes, such as the aerobic oxidation of Fe^{II} or the anaerobic reduction of elemental S^0. Results from this pilot project will be beneficial to a larger metagenomic sequencing effort currently underway (DOE-JGI), where 20 geochemically-diverse geothermal sites have been included for metagenomic sequencing. Results from the pilot project demonstrate that these environments offer significant advantages for interpreting metagenomic data and for employing genomic resources to address specific hypotheses in microbial community ecology, population biology, microbial evolution and geobiology.

The Utility of EST Libraries as a Clone Resource and a Tool for Improvement of Gene Annotation

Scott E. Baker[1] (scott.baker@pnl.gov), **Frank R. Collart**[2]* (fcollart@anl.gov), Sarah Zerbs,[2] and Jessica Saunders[2]

[1]Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, Washington, and [2]Biosciences Division, Argonne National Laboratory, Lemont, Illinois

For eukaryotic organisms, the complexity of the gene structure can often provide a challenge to obtaining a resource for the cloning process. As part of the process to improve gene models, genome sequencing projects for eukaryotic organisms often utilize cDNA libraries generated from the target organism. For some organisms these libraries are available in high density plate formats but often are uncharacterized. We analyzed sequence data from two EST libraries (*Aspergillus niger & Laccaria bicolor*) generated at the U.S. DOE Joint Genome Institute. These libraries contain individually arrayed clones and our analysis indicates that a substantial number of clones are full length or nearly full length. In the case of the *A. niger* EST library, ~2400 unique full length or nearly full length clones were identified. These clones are a valuable resource in that they provide a cost effective and quick route to an expression clone. A subset of glycosyl hydrolase EST clones from the *A. niger* library was resequenced to validate the integrity of the clones and then used for expression in bacterial and yeast systems. The analysis process also identified a subset of clones that were not consistent with the gene model. This more intensive analysis of these EST library resources provides opportunities to improve gene models for these organisms and material for downstream expression and characterization experiments linked to the gene annotation.

The Genome of the Eukaryote *Phaeocystis antarctica*: A Dominant Phytoplankter and Ice Alga in the Southern Ocean

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

Stanford University, Stanford, California

Phaeocystis is an important unicellular eukaryotic photosynthetic phytoplankton genus distributed throughout the world's oceans, with representatives living in the open ocean as well as in sea ice. A unique attribute of *Phaeocystis* is its ability to form a floating colony with hundreds of cells embedded in a polysaccharide gel matrix that can multiply massively during blooms. The *Phaeocystis* genus contributes approximately 10% of annual global marine primary production, equivalent to 4 billion metric tons of carbon a year. Given that up to 50% of this carbon is in the form of polysaccharides, 2 billion tons of polysaccharide gel-carbon is produced by *Phaeocystis* sp. each year. This is a key process in the marine carbon cycle and it contributes to the higher efficiency of atmospheric CO_2 drawdown by this genus compared with other phytoplankton.

Phaeocystis is also an important producer of 3-dimethylsulphoniopropionate (DMSP), the precursor of the climate gas dimethylsulphide (DMS). Biogenic DMS contributes about $1.5x10^{13}$ g sulfur to the atmosphere annually and plays a major part in the global sulfur cycle, in cloud formation, and potentially in climate regulation. Sequencing the genome of

P. antarctica, a major bloom-forming organism and ice alga endemic to the Southern Ocean, will greatly improve our understanding of its physiology and its role in the Antarctic ecosystem where climate change is manifesting itself faster than in any other oceanic region. Forecasting the future distributions of *P. antarctica* as global warming continues and ice cover diminishes may play a critical role in predicting possible variations in CO_2 draw-down and particle sedimentation in this region. The sequence information will have a tremendous impact on the research community, providing researchers with genomic tools that will enable them to determine the potential metabolic pathways within the organism and the ways in which these pathways respond to fluctuations in temperature, salinity, light, and nutrients in the environment (e.g. through the development and use of microarray technology and qPCR strategies). More specifically, several groups spanning chemical, atmospheric and biological research will use the sequence information to investigate the regulation of carbon fixation, polysaccharide excretion, DMS production, stress-response and bloom demise, cold acclimation and antifreeze protein production. The construction of a full genome array will help identify many of the genes involved in these processes.

Small RNA Analysis of the Hyperthermophile Genus *Pyrobaculum*

David L. Bernick* (dbernick@soe.ucsc.edu), Matthias Hoechsmann, and Todd M. Lowe

Biomolecular Engineering, University of California, Santa Cruz, California

We have recently completed a high-throughput analysis of small RNA expression across four species in the crenarchaeal genus *Pyrobaculum*, in collaboration with the U.S. DOE Joint Genome Institute. The expression data, derived from 454/Roche sequencing of short RNAs, is a first look at the “small transcriptomes” of three new genomes decoded just last year as part of the DOE Community Sequencing Program. The small RNA sequencing data has provided a wealth of evidence advancing research in multiple areas: details of CRISPR expression and processing, cross-species conservation of expressed antisense elements, novel non-coding RNAs, RNA secondary structure information, 5' transcription boundaries supporting a novel instance of translational programmed -1 frameshift regulation, validation of previously predicted sRNA genes, and relative sRNA expression levels which are more variable than expected.

The *Pyrobaculum* species analyzed encode between five and seven CRISPR arrays containing between two and up to eighty embedded sequence elements (spacers), punctuated by direct repeat (DR) sequences. CRISPR array expression appears consistent with the model of a single leader-resident promoter per array, incrementally terminated at each embedded DR, and subsequently processed to yield individual spacer sequences. Spacer expression appears to decrease in a non-linear fashion with distance from the promoter. Notable exceptions exist where spacers are either absent or surprisingly highly-expressed.

Cis-antisense expression was found overlapping transposases, unclassified but highly repetitive “Rep” genes (unique to *Pyrobaculum*) and putative biosynthetic genes. These small transcripts range in size across the full spectrum of RNAs purified for sequencing, 20-70 nucleotides. The antisense transcripts represent a potential method for gene regulation in these species, and likely other archaea.

sRNAs are heavily expressed within this group. The transcription boundaries align well with current computational predictions, validating the efficacy of these methods. Novel transcripts, consistent with new sRNAs between known genes, are also present. These

experimental results will likely provide new training data and improved accuracy for these predictive methods.

Within the *Pyrobaculum* group, we find expression of a suspected RNA gene with supporting compensatory mutations indicative of RNA secondary structure. It appears that spontaneous hydrolysis in loop regions has yielded secondary structure data consistent with our computational structure predictions. Potentially, this method can be extended using enzymatic or chemical means to provide genome-wide analysis of RNA secondary structure.

Transcription of mRNA molecules is present at low levels in this data set. In one instance, multiple transcripts support a model for -1 frameshift regulation of the carbamoyl-phosphate synthase gene. This gene is regulated via translational means in other model organisms, and the *Pyrobaculum* clade appears to add a new variant.

High throughput sequencing and genome-scale visualization have provided a mechanism where we can quickly produce multiple-genome expression data with sequence-level resolution rapidly and effectively. In this pilot study, we received 213,000 sequence reads from a single Roche/454 sequencing plate. These were aligned to their respective genomes and made available for internal inspection on the UCSC Archaeal Genome browser within a few hours of receiving the raw data from JGI. We anticipate this data and further analyses will be publicly available early this summer.

Generation of an Insertional Mutant Library in *Brachypodium distachyon*

Jennifer Bragg[1]* (jbragg@pw.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, California

To facilitate the use of *Brachypodium distachyon* as a model organism for cereal and energy crops, researchers have initiated construction of a suite of genomic resources. The most fundamental component of this collection is the genome sequence that is currently available at preliminary 4X coverage, with the final 8X assembly anticipated to be completed by the middle of 2008. The emerging collection of tools also includes multiple BAC libraries, SNP markers, physical and genetic maps, and germplasm resources. To complement these efforts, we are generating a population of sequence-indexed insertional mutants using a combination of *Agrobacterium*-mediated T-DNA tagging and transposon-mediated mutagenesis. Within this collection we plan to generate insertional mutant knock-out lines for study of individual genes of interest, promoter-trapping lines for identification of sequences that will be useful for expression of transgenes in *B. distachyon* and other grasses, and activation-tagged lines that are suited for analyses of genes with redundant functions for which knockout lines fail to produce a phenotype. The project objectives target generation of >7,500 insertional mutants, sequencing >6,000 insertion sites, annotation using the genomic sequence to identify the genes affected, and compilation of the data into a searchable website to provide researchers with a means to order lines with mutations in regions of interest. As a result, this project will provide a large, freely available collection of sequence-indexed mutants to researchers studying grasses and grains. Currently, we are evaluating different media, selectable markers, and

promoters in order to optimize transformation efficiency. Here we describe our procedures and the current status of our project.

The *Porphyra* Genome: Promoting Resource Development and Integrative Research in Algal Genomics

Susan H. Brawley[1]* (brawley@maine.edu), John W. Stiller[2] (stillerj@ecu.edu), Elisabeth Gantt,[3] and Arthur C. Grossman[4]

[1]University of Maine, Bangor, Maine; [2]East Carolina University, Greenville, North Carolina; [3]University of Maryland, College Park, Maryland; and [4]Carnegie Institution of Washington, Stanford University, Stanford, California

The red alga *Porphyra purpurea* (Bangiophyceae) is a new JGI project, and we are now preparing material for genomic sequencing. *Porphyra* has a well-characterized biphasic sexual life history, an ancient fossil record, and is an important human food (laver or nori). The National Science Foundation's Division of Integrative Organismal Systems recently recommended an associated project for funding ("Research Coordination Network: The *Porphyra* genome: promoting resource development and integrative research in algal genomics"). Assuming that NSF's Division of Grants and Agreements accepts the recommendation and makes the award, this RCN (2008-2013) will provide opportunities to a large group of scientists directly involved in analyses of the *Porphyra* genome, and will encourage collaborations with researchers from other algal genomics projects. We offer this poster as an announcement of these potential opportunities; please let us know if you are interested. The RCN will meet annually and will include training and discussions on comparative and experimental approaches with the genomic data that will initiate further research projects and proposals; it will include both graduate students and postdoctoral fellows. We anticipate that RCN-aided analysis of the *Porphyra* genome will help to advance biological knowledge in a wide range of areas: for example, it will lead to new understandings of chloroplast evolution, nuclear-organelle co-evolution, gene transfers between the chloroplast and nucleus and import of proteins into the chloroplast, advance studies of nitrogen and carbon metabolism in photosynthetic organisms, foster comparative investigations of how multicellular complexity arose independently in different eukaryotic groups, and provide a wealth of new markers for studies in molecular evolution and systematics. The JGI and potential NSF RCN projects will develop new collaborations around central questions that impact many key societal needs, including multitrophic aquaculture (e.g., bioremediation) and integrated analyses of photosynthetic and photoprotective pigments as they affect carbon metabolism on our warming planet.

Genome Sequencing of Budding and Non-Budding Stalked Bacteria from Aquatic Environments

Pamela J.B. Brown[1]* (pjbonner@indiana.edu), Jeong-Hyeon Choi,[2] Kwangmin Choi,[3] Ankita Bhan,[3] Ellen N. Weinzapfel,[1] Sun Kim,[2,3] and Yves V. Brun[1] (ybrun@indiana.edu)

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

The constellation of shapes and sizes among bacteria is as remarkable as it is mysterious. Why should some bacterial species adopt such diverse shapes as a bedspring coil, a star or a partly eaten donut? No one really knows. However, the precise reproduction and

evolutionary conservation of these shapes indicate that they play an important role in the life of bacteria. Despite recent progress in understanding the mechanisms that control cell shape determination, or morphogenesis, we still do not understand how bacterial cells generate specific shapes, or what the function of bacterial morphological changes is. This project focuses on a group of bacteria, the prosthecate or stalked bacteria, which provides a well-defined and simple example of morphological change. The stalks synthesized by members of the □-Proteobacteria take up diffuse compounds from water sources, a feature that could be exploited for bioremediation, specifically the uptake of toxic compounds from contaminated water sources. Furthermore, extracellular polysaccharides from some of the stalked bacteria sequester metals, a feature that could be used to remediate environments impacted by metal toxicity. In this work, five non-budding prosthecate bacteria (*Asticcacaulis biprosthecum, Asticcacaulis excentricus, Brevundimonas subvibrioides, Prosthecobacter vanneervenii* and *Stella humosa*), four budding prosthecate bacteria (*Ancalomicrobium adetum, Hyphomicrobium denitrificans, Prosthcomicrobium hirschii* and *Rhodomicrobium vannielii*), and two non-budding, non-stalked bacteria (*Brevundimonas diminuta* and *Caulobacter segnis*) have been selected for genome sequencing. Our goal is to produce a high quality draft sequence of these genomes. The genome sequences will be used to generate sets of gene predictions, annotate the sequences, and perform comparative genomics of these and closely related genomes of stalked and non-stalked bacteria. The genomes will be analyzed with respect to mechanisms for the biosynthesis and function of stalks, the biosynthesis and regulation of extracellular polysaccharide, the extent of conservation of regulatory pathways for stalk and adhesin biosynthesis, and the interesting and potentially useful physiological properties of these organisms. Comparative analysis of the genome sequences of the stalked bacteria in this proposal will further our understanding of prokaryotic stalk synthesis and stalk protein targeting and provide the information necessary to engineer stalked bacteria to efficiently remove toxic compounds from contaminated water sites.

The Genome of *Dechloromonas aromatica* strain RCB and Redundancy in the Genetic Pathway for Aerobic Hydrocarbon Oxidation

Kathryne G. Byrne-Bailey[1]* (KByrne@nature.berkeley.edu), Forest M. Kaiser,[1] Kennan Kellaris Salinero,[1] Nandini Krishnamurthy,[1] Jake Gunn-Glanville,[1] Yvonne Sun,[1] Romy Chakraborty,[1] Kimmen Sjolander,[1] William S. Feil,[1] Helene Feil,[1] Eric Alm,[2] Katharine Huang,[2] Morgan Price,[3] Keith Keller,[3] Adam Arkin,[3] Alla Lapidus,[4] Stephan Trong,[4] Genevieve Di Bartolo,[1] Frank W. Larimer,[5] Paul M. Richardson,[4] Laurie A. Achenbach,[6] and John D. Coates[1]

[1]University of California, Berkeley, California; [2]Massachusetts Institute of Technology, Boston, Massachusetts; [3]Virtual Institute of Microbial Stress and Survival, Berkeley, California; [4]DOE Joint Genome Institute, Walnut Creek, California; [5]Oak Ridge National Laboratory, Oak Ridge, Tennessee; and [6]Department of Microbiology, Southern Illinois University, Carbondale, Illinois

The facultative beta proteobacterium*, Dechloromonas aromatica* strain RCB is unique in its ability to aerobically and anaerobically oxidize benzene and other aromatic hydrocarbons. Its complete genome is a single circular chromosome of 4,501,104 base pairs (bp) with an average G+C content of approximately 60 %. Annotation revealed 4280 putative coding genes of which 69.32 % were assigned a putative function.

The genome revealed that RCB is well adapted for survival in the environment. This adaptation includes the presence of a large number of signaling proteins in relation to its genome size. The genes encoding these proteins were apparently recently duplicated and a high proportion were homologous to two component histidine kinase systems. Also predicted to enable survival in the environment were the enzymes involved in the Wood-Ljungdahl pathways, the Calvin cycle and three systems involved in energy and carbon capture via pyruvate conversion to actyl-CoA, suggesting flexibility in carbon utilization and energy sources.

D. aromatica strain RCB has previously been reported to aerobically degrade a number of aromatic hydrocarbons, in addition to their anaerobic oxidation coupled to Fe(II), nitrate or per(chlorate) reduction. Annotation revealed redundancy in the genes involved in the aerobic oxidation pathways of benzene and toluene. In total a putative toluene-4-monooxygenase, phenol hydroxylase, benzoate-1,2-monooxygenase and three catechol-2,3-dioxygenase genes were identified within a 20 kb region of the chromosome. Multiple clusters of putative genes for aerobic phenolic compound degradation were also annotated in the same genomic segment. Strain RCB is only one of five bacteria with more than one BMM (bacterial multi-component monooxygenase). The “toluene-4-monooxygenase” was encoded by a six gene cluster, with high similarity to a gene cluster in *Azoarcus* sp. EbN1. The cluster was flanked on the 3' side by a number of signal transduction histidine kinases indicative of environmental sensing and regulation and on the 5' end by a IS4 transposase. Strain RCB also encoded several proteins related to toluene resistance.

Reverse transcription (RT)-PCR analysis has been performed to elucidate whether the apparent gene redundancy in the aerobic pathways for the oxidation of benzene and toluene was observed at the transcription level with different substrates. The three catechol-2,3-dioxygenase genes were also further investigated for expression and pathway involvement.

These studies demonstrate the utilization of this genome sequence, the sensitivity of this organism to its environment and its ability to customize and optimize hydrocarbon catabolism based on some as yet to be defined criteria.

Community-Involved Genome Annotation and Analysis at JGI-LANL: Facilitating Publication of Genome Papers Through Bioinformatics Support and Training

Jean Challacombe, **Gary Xie*** (xie@lanl.gov), Diego Martinez, Ravi Barabote, Monica Misra, and Thomas Brettin

Los Alamos National Laboratory, Los Alamos, New Mexico

The role of the genome annotation and analysis team at JGI-LANL is to facilitate publication of JGI genome papers and provide bioinformatics support and training to promote community-involved genome annotation and analysis. Since March 2007, we have hosted 9 JGI collaborators as part of our genome explorer seminar series. In projects where JGI-LANL team members played a leading role in the analysis and preparation of genome papers, 6 genome papers have been published, 1 book chapter has been in press and 5 are submitted or near submission. In addition to our microbial genome effort, our eukaryotic genome annotation team is working with the annotation team at JGI-PGF and fungal research community to provide high quality manually curated annotations of fungal genomic sequences. We have hosted 1 off-site annotation Jamborees to promote

community involved analysis for publishing a full analysis and annotation of *Postia placenta*, *Trichoderma virens* and *Trichoderma atroviride* chromosomes. Work is under way to publish its genome papers.

Automated Accurate, Concise and Consistent Product Description Assignment for Microbial Transporter Proteins

Yun-juan Chang* (yjs@ornl.gov), Miriam Land, Frank Larimer, and Loren J. Hauser (hauserlj@ornl.gov)

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge Tennessee

Since the advent of genomic sequencing technologies, an immense amount of genomic data has been generated. The complete sequence of over 700 microbial genomes has been published to date. New technology advancement will greatly accelerate the increase rate of sequenced genomes. Automated and effective computational methods are required to analyze the ever-growing sequence information.

Cell membrane transport systems play essential roles in cellular metabolism and activities. An in-depth evaluation of transporter proteins is critical to the understanding of physiology of the sequenced organism. However, it is often problematic to annotate these proteins by current methods due to large and complex transporter gene families and the multiple transporter paralogs in many organisms. We here present our work aimed at a genome wide, systematic, and automated annotation of transporters.

A comprehensive analysis of nearly 40000 identified transporters was performed using Interpro, COGs, Pfam, tfam and TCDB (Transporter classification database), as well as TMHMM predictions. A rule based annotation system combining the output from searches against those databases/tools and domain architecture has been developed in order to provide accurate, concise, and consistent product descriptions for most microbial transporters. This multidimensional approach increases both the accuracy and sensitivity when compared to the use of single system such as Transporter database (TransportDB) or TCDB searches. For each annotated microbial genome, the system provides a list of all the transporters, the product description that will be assigned, and all other information used in their identification. In addition, it will provide a numerical synopsis of the transporter organized by TCDB transporter classification (TC) number scheme.

The above tool has been demonstrated to cover more than 84% transporters as tested with selected genomes. This system will be incorporated into the Oak Ridge National Laboratory's automated annotation pipeline as an additional feature.

Molecular Identification of Putative Reductive Dehalogenase in Novel *Dehalococcoides* Species

Wai Ling Chow, Dan Cheng, and **Jianzhong He*** (jianzhong.he@nus.edu.sg)

Division of Environmental Science and Engineering, National University of Singapore, Singapore

An enrichment culture MB containing *Dehalococcoides* species could reductively dechlorinate PCE/ TCE to *trans*-DCE as their predominant end product. The reductive dehalogenases (RDases) responsible for the generation of *trans*-DCE has yet to be identified. Therefore, degenerate primers targeting conserved regions in RDases were used to amplify putative reductase genes from culture MB. Five unique RDase gene fragments (approximately 1.7kb) were identified, and transcriptional studies of PCE grown MB cultures suggested that cDNA5 was involved in the generation of *trans*-DCE. Sequence comparisons of cDNA5 to proteins in the public databases revealed weak homology. The cDNA5 protein sequence was found to exhibit the characteristics of RDases: the twin arginine motif (RRXFXK) in the form of RRDMFK at the N terminal and the presence of an eight-iron ferredoxin cluster binding motif $(CXXCXXCXXXCP)_2$ at the C terminal. Together with the identified putative cDNA5, a small ORF was found to cotranscribe with cDNA5 and is proposed to be involved in the membrane association of this RDase. The cDNA5 gene was absent in *Dehalococcoides* mixed cultures that do not generate *trans*-DCE but was detected in six out of eleven *trans*-DCE producing mixed cultures. As similarity of 16S rRNA gene sequences in *Dehalococcoides* populations does not necessarily translate into analogous dehalogenating ability, the identification of RDase involved in *trans*-DCE production will greatly complement the current inadequate 16S rRNA gene approaches in assessing and monitoring in situ PCE/TCE dechlorination at chloroethene-contaminated sites.

The JGI Community Sequencing Project for Conifer ESTs

Jeffrey F.D. Dean* (jeffdean@uga.edu), Glenn T. Howe, Kathleen Jermstad, David B. Neale, and Deborah L. Rogers

University of Georgia, Athens, Georgia

Conifers constitute an ancient and diverse branch in higher plant evolution. Some conifer species dominate modern day ecosystems that are repositories for large amounts of terrestrial sequestered carbon, while others exist in populations numbering tens of individuals. Conifer forests are among the most productive in terms of annual lignocellulosic biomass generation and coniferous trees are the preferred feedstock for much of the forest products industry, one of the most energy-intensive manufacturing sectors of the U.S. economy. Breeding programs to improve conifers have been in existence for more than 50 years, but progress has been slow due to the large size of the trees and their generally slow progress to sexual maturity. Climate change and exotic forest pests are threatening certain conifer populations, but a general dearth of genomic resources and tools limit our capacity to address many of these issues and problems.

The U.S. DOE Joint Genome Institute, through its Community Sequencing Program, has initiated a project that will great increase the publicly available EST sequences for loblolly pine and a wide variety of other conifer species, including representative species from all families of the Coniferales. Targets for the project include paired-end reads from 250,000 cDNAs using Sanger sequencing, and several million reads using the Roche/454

pyrosequencing platform. This presentation will provide an update on the current status of the project, and specific targets and species will be discussed. Discussions will also include ways in which this project will interact with other ongoing conifer and gymnosperm genome projects.

Accurate, Comprehensive Binning of Deeply Sampled Community Genomic Sequence Reveals Unexpected Diversity and Allows Reconstruction of Low-Abundance Bacterial Genomes

Gregory J. Dick[1]* (gdick@berkeley.edu), Anders Andersson,[1] and Jillian F. Banfield[1]

[1]Department of Earth and Planetary Science, University of California, Berkeley, California

A major challenge facing sequence-based community genomics is assignment of metagenomic fragments to microbial species, populations, or groups at some taxonomic level. Known as 'binning', this is a prerequisite for realizing some of the most valuable opportunities that metagenomics offers, including assignment of ecological and biogeochemical functions to particular organisms or groups, and assessment of community structure and population level genomic diversity. Particularly daunting is metagenomic analysis of low-abundance community members. Despite accounting for only a very minor fraction of cells or biomass, these organisms often play critical ecological roles and thus represent keystone community members.

Here we present binning analysis of community genomic sequence from biofilms inhabiting acid mine drainage (AMD) in the Richmond Mine at Iron Mountain, CA. The biofilms are dominated by chemolithoautotrophic organisms that are sustained by the oxidation of Fe(II). This microbial activity drives the dissolution of pyrite and the generation of AMD, a worldwide problem of significant environmental concern. We used tetranucleotide frequency and self-organizing maps (SOMs) to evaluate a dataset of AMD community genomic sequence that included both previously assembled/identified sequences as well as unassigned sequence fragments. The tetra-SOM effectively resolved the vast majority of assembled genomic sequence, including even small genomic fragments (1.5-kb). Previously unrecognized regions of tetranucleotide frequency signature were revealed, corresponding to novel low-abundance organisms and putatively extra-chromosomal elements of dominant organisms (i.e. phage or plasmid). Reconstruction of significant portions of genomes from low-abundance community members was possible, including novel acidophilic *Actinobacteria*. The ability to resolve genomes was a function of phylogenetic relatedness rather than G+C content: distantly related genomes with identical G+C content were effectively resolved whereas more closely related genomes with distinct G+C content showed some regions of overlap. Overall, our results demonstrate that tetra-SOM is a valuable method of binning and visualizing community genomic data.

MyJGI: Tools for Collaborators

Joni Fazo[1]* (jbfazo@lbl.gov), Annette Greiner,[2] Anthony Kosky,[2] Rene Perrier,[2] David Pletcher,[1] Kristen Taylor,[1] and Arkady Voloshin[2]

[1]DOE Joint Genome Institute, Lawrence Livermore National Laboratory, Livermore, California, and [2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California

The JGI Informatics department launched http://my.jgi.doe.gov, aka MyJGI, in October of 2007. This new web site provides information and tools useful to collaborators with sequencing projects at JGI.

The site currently features three main areas:

- General Info: http://my.jgi.doe.gov/general
- Sample Info: https://my.jgi.doe.gov/csi
- Status Reporter: https://www.jgi.doe.gov/collaborators/status.php

The General Info area provides useful information to help users get started on a project with JGI, recommended protocols, JGI policies, and other resources. The Getting Started page offers step-by-step instructions for initiating a project. The protocols offer successful examples for preparing samples or libraries to be sent to the JGI. JGI's policies on project scheduling, finishing, publication, and data release are detailed in this area, and a sample user agreement is available for download.

The Sample Info area provides access to CSI – the Collaborator Sample Information tool. Prior to shipment of samples to the JGI, collaborators use CSI to provide project managers and scientists with information, such as gel image, volume and weight, required to get samples into the JGI pipeline efficiently. JGI project managers then use internal web forms to review and approve each sample for shipment. The process integrates with downstream systems such as the JGI Global Project Management Tracking System (GPTS) and LIMS. The Scientific and Institutional Applications Group (S&IA) created this system using the web technology Ruby on Rails. Data is stored in a MySQL database; the system also interfaces with the JGI GPTS Oracle based database system. In the future, we plan to offer increased integration between CSI and GPTS, as well as support for metagenomic projects and projects focused on new technologies.

The Status Reporter area provides collaborators with real-time information about the status of each of their sequencing projects. For most projects, the status report includes detailed information about the planned lifecycle of the project as well as release dates and other milestones. We have recently added full support for eukaryote projects, and all active projects now show at least basic sequencing status information. Once a collaborator logs in, they are presented with a list of their own proposals, from which they can select one to view status information. A single click brings up the status information for all projects that are part of that proposal. If the collaborator has only one proposal, the project view pops up immediately after login. The CSR was also created by the S&IA group, and was written in PHP with dynamic use of CSS.

Note: The Sample Info and Status Reporter areas are available only to current collaborators and require a user name and password to access. Collaborators may request a login from their JGI project manager. We have recently implemented single sign-on for MyJGI, so that a user who has logged in to use either the Sample Info area or the Status Reporter can use the other without logging in again.

Genetic Noise Control via Protein Dimerization

Cheol-Min Ghim* (cmghim@llnl.gov) and Eivind Almaas

Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California

Gene expression in a cell entails random reaction events occurring over disparate time scales. This molecular noise, often resulting in phenotypic and population-dynamic consequences, sets a fundamental limit to biochemical signaling. There have been numerous studies correlating the structure of cellular reaction networks with noise tolerance, but only a few efforts have been made to understand the dynamical role of protein-protein association. Ample examples of homo- or hetero-oligomeric proteins may provide a rationale for genetic noise control. With this motivation we have developed fully stochastic models of simple autoregulatory gene circuits, integrating the quantitative results of previous *in vivo* and *in vitro* studies. In particular, we explicitly consider the fast binding/unbinding kinetics among proteins, RNA polymerases, and promoter/operator sequences of DNA. In the presence of protein dimerization, both the monomer and dimer copy numbers show significantly reduced fluctuations. The frequency content of noise power spectra is dramatically shifted to the physiologically irrelevant high-frequency regime. This behavior persists throughout the model systems, regardless of the sign of regulation, binding affinities or rate parameters within moderate variations from typical values. Specific binding of regulatory proteins provides a buffer that mitigates erratic eruption or depletion of genetic activities. The capacity of the buffer is a non-monotonic function of the association-dissociation rates. Because the protein oligomerization *per se* does not require extra protein components, it lends a basis for the rapid control of stochastic fluctuation in response to changing environment, and may also assist the rational design of proteins or synthetic gene circuits for engineering purposes.

An Experimental Approach to Map Ligands with Binding Proteins and Improve Gene Annotation

Sarah E. Giuliani, **Frank R. Collart*** (fcollart@anl.gov), and Ashley M. Frank

Biosciences Division, Argonne National Laboratory, Lemont, Illinois

We evaluated a fluorescent thermal shift assay (TSA) as a high throughput approach to improve gene/protein functional assignments. The assay involves monitoring changes in the fluorescence signal of SYPRO orange as it interacts with a protein undergoing thermal unfolding. Protein unfolding with increasing temperature often results in exposure of hydrophobic regions and eventual protein aggregation. Ligand binding to a target protein can stabilize a protein's native state reflected in the increased melting temperature (Tm) of the bound protein. For preliminary evaluation of the utility of the Thermal Shift Assay, we elected to focus on periplasmic binding proteins and specifically those annotated as part of an ABC transporter system. We cloned, expressed and purified a set of 10 periplasmic amino acid-binding proteins for TSA screening against known ligands. This set included a set of positive control targets having PDB models of proteins bound to native ligand as well as homologs and uncharacterized proteins from *Escherichia coli*, *Shewanella oneidensis Campylobacter jejuni*, and *Salmonella typhimurium*. Our assay validated the binding specificity of the known proteins and was able to unambiguously identify preferential binding for homologs and proteins of uncharacterized binding specificity. Relative binding affinities for specific ligands could be determined from differential Tm

shifts using a series of increasing ligand concentrations. This set of targets show high selectivity for specific amino acids in the ligand library with temperature shifts or 6-12°C observed for specific ligands. The assay is currently being expanded to analyze the set of branched chain amino acid binding proteins from *Rhodopseudomonas palustris.*

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357.

Brachypodium Genome and its Syntenic Relationship with Cereal Crops

Yong Qiang Gu[1]* (Yong.Gu@ars.usda.gov), Naxin Huo,[1] John P. Vogel,[1] Gerard R. Lazo,[1] Frank M. You,[1] Olin D. Anderson,[1] Yaqin Ma,[2] Jan Dvorak,[2] and Ming-Cheng Luo[2]

[1]USDA-ARS, Western Regional Research Center, Genomics and Gene Discovery Unit, Albany, California, and [2]Department of Plant Sciences, University of California, Davis, California

Because of its small genome size (~350 Mb) and several desirable attributes, *Brachypodium distachyon* is emerging as a model system for temperate grasses, including important crops like wheat and barley. Analysis of 10.9% of the *Brachypodium* genome based on 64,696 BAC end sequences (BES) revealed that the genome consists of ~ 18.4 % of repetitive elements (TEs), with 11% of known TEs and 7.4% of unique *Brachypodium* TEs. Sequence analysis indicated that approximately 21.2% of the *Brachypodium* genome represents coding sequence. The BESs were integrated into the BAC-based physical maps of the *Brachypodium* genome, which allows for comparison of gene order and contents with that of the rice genome at a genome-wide level. Large conserved genomic regions were readily identified between the two small grass genomes. We also analyze the sequence conservation at the microcolinearity level by comparing sequenced *Brachypodium* BACs with the orthologous regions from rice. Genomic rearrangements, differential gene amplification and deletion appeared to be the common evolutionary events that caused variations of microcolinearity at different orthologous genomic regions. In addition, several annotated genes in *Brachypodium* BACs have matches to the wheat deletion bin mapped ESTs. In some cases, genes in the same BACs matched to the wheat ESTs that were mapped to the same deletion bins, suggesting that the *Brachypodium* genome will provide useful information in placing the order of mapped wheat ESTs within the deletion bins and developing specific markers in the targeted regions of wheat chromosomes.

Whole-Genome Sequencing *Arabidopsis lyrata* and *Capsella rubella*: Two Close Relatives of *A. thaliana*

Ya-Long Guo[1]* (ya-long.guo@tuebingen.mpg.de), Jan-Fang Cheng,[2] Richard M. Clark,[1] Christa Lanz,[1] Korbinian Schneeberger,[1] Stephan Ossowski,[1] Norman Warthmann,[1] Joy M. Bergelson,[3] Justin O. Borevitz,[3] Brandon S. Gaut,[4] Anne E. Hall,[5] Charles H. Langley,[6] Barbara Neuffer,[7] June B. Nasrallah,[8] Klaus F. X. Mayer,[9] Magnus Nordborg,[10] Outi Savolainen,[11] Yves Van de Peer,[12] Stephen I. Wright,[13] Jeremy Schmutz,[14] Dan Rokhsar,[2] and Detlef Weigel[1] (weigel@tuebingen.mpg.de)

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany; [2]DOE-Joint Genome Institute, Walnut Creek, California; [3]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois; [4]Department of Ecology and Evolutionary Biology, University of California, Irvine, California; [5]Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, Illionis; [6]Section of Evolution and Ecology, University of California, Davis, California; [7]Department of Systematic Botany, Universität Osnabrück, Osnabrück, Germany; [8]Department of Plant Biology, Cornell University, Ithaca, New York; [9]Munich Information Center for Protein Sequences, Institute for Bioinformatics, Gesellschaft fur Strahlenforschung Research Center for Environment and Health, Neuherberg, Germany; [10]Molecular and Computational Biology, University of Southern California, Los Angeles, California; [11]Department of Biology, University of Oulu, Oulu, Finland; [12]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Ghent, Belgium; [13]Department of Biology, York University, Toronto, Canada; and [14]Stanford Human Genome Center, Stanford University, Stanford, California

Arabidopsis thaliana was the first plant, and the third multicellular organism after *Caenorhabditis elegans* and *Drosophila melanogaster* to have a reference strain completely sequenced. In order to leverage the extensive genomic information available for *A. thaliana*, including an increasing amount of knowledge of within-species diversity, as well as to better understand plant genome evolution in general, we proposed two close relatives for genome sequencing (*Arabidopsis lyrata* and *Capsella rubella*). *A. lyrata* is the closest well-characterized relative in the same genus as *A. thaliana*, and *Capsella* is the closest well-characterized genus. Both genomes are being ~8-fold shotgun sequenced. An assembly for *A. lyrata* is being finalized, while sequencing of *Capsella rubella* is in progress. Comparing the genomes of *A. thaliana*, *A. lyrata*, and *C. rubella* provides an unprecedented opportunity to understand key aspects of plant genome evolution in species that share high enough sequence identity that intermediate genomic changes can be directly determined. In detail, assignment of ancestral states for *Arabidopsis* using *C. rubella* as outgroup; genome-wide analysis of microstructural evolution in plants; elucidation of genome size variation in closely related species; functional information from these species will likely be directly transferable to other relevant plant species, certainly to the many heavy-metal accumulators in the same family (e.g., *Arabidopsis halleri*), which are of interest for bioremediation; much of functional information from these species will also apply to the fairly closely related poplar, of interest for carbon sequestration and energy production, and already sequenced by the JGI.

A High-Density SNP Map in *Brachypodium distachyon*

Naxin Huo[1, 2]* (nhuo@pw.usda.gov), Stephanie McMahon,[1] Frank M. You,[1,2] Gerard R. Lazo,[1] Mingcheng Luo,[2] Yong Q. Gu,[1] Olin D. Anderson,[1] David Garvin,[3] and John Vogel[1]

[1]Genomics and Gene Discovery Research Unit, USDA-ARS, Western Regional Research Center, Albany, California; [2]Department of Plant Sciences, University of California, Davis, California; and [3]USDA-ARS Plant Science Research Unit, University of Minnesota, St. Paul, Minnesota

To fully realize the potential of the small grass *Brachypodium distachyon* as a model system it is necessary to develop a high-density genetic linkage map. To achieve this, we have initiated a project to map ~1,000 single nucleotide polymorphism (SNP) markers using the Illumina genotyping platform. This level of resolution should result in a marker every 250-400 Kb assuming a genome size of 300 Mbp. To maximize the utility of this map, marker development was initially targeted to produce markers evenly spaced along the physical contigs. Since the completion of the 4x draft sequence by JGI, we are now targeting marker development to produce markers evenly spaced along the 4X scaffolds. To date, unique BAC end sequences derived from line Bd21 have been used to design 3,483 primer pairs and amplify the corresponding DNA from line Bd3-1. By sequencing the amplicons, we identified 1,975 SNPs at 632 loci. The SNP markers were distributed on 329 physical contigs and 29 super contigs, which cover about 280 Mb of the *Brachypodium* genome. The first 571 BAC end sequences corresponding to the marker loci were Blasted against the Rice whole genome (TIGR Pseudomolecule assembly release 4) and 320 out of the 571 loci (56%) had homologous sequences in the rice genome. An F_2 population of ~480 individuals derived from crosses between inbred lines Bd21 and Bd3-1 will be used to construct the genetic linkage map. We plan to complete the SNP map in time to aid and verify the final assembly of the 8x shotgun sequence produced by JGI.

Prodigal: A New Prokaryotic Gene Identification Program with Enhanced Translation Initiation Site (TIS) Prediction

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

Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

High-quality annotation of microbial genomes remains an ongoing challenge for the U.S. DOE Joint Genome Institute. For the past several years, Oak Ridge National Laboratory (ORNL) has aided in gene prediction and functional analysis of numerous microbial organisms. As a result of the comprehensive review and curation of a large number of genomes ranging from low-GC to high-GC and from bacteria to archaea, numerous areas of improvement have been found. Predicting the correct number of genes, reducing the number of false positives, correctly locating short and laterally transferred genes, performing robustly in high-GC-content genomes, and accurately finding the translation initiation site (TIS) of genes continue to be challenges that will enhance the quality of the final JGI annotations submitted to Genbank and placed in IMG.

Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) was developed at ORNL to address many of the “real-world” challenges discovered through many hours of manual curation of microbial genomes. In particular, the previous pipeline (based on the genefinders Critica and Glimmer) often lengthened genes in high-GC content genomes and incorrectly omitted genes that would overlap the erroneously long genes. We decided to

address this issue with a new gene identification algorithm that would perform robustly in high-GC content genomes. Prodigal's self-training methodology is based on a detailed analysis of the GC-frame-plot of the organism in question. The training process consists of determining the statistical significance of G and C in different frame positions and performing a dynamic programming algorithm using this information to construct an initial training set of genes. This is a novel approach compared to other programs, which construct their training sets based merely on all open reading frames (ORFs) above a particular length. Our implementation of a coding scoring function based on this training set was found to perform well in both low-GC and high-GC genomes.

The other improvement to the annotation pipeline is improved start site prediction. Prodigal contains a novel method for examining the upstream regions for ribosomal binding site (Shine-Dalgarno) motifs. The statistical significance of various motifs relative to the background is determined automatically by an iterative algorithm which learns the organism's preference for various RBS motifs. Results for this enhanced start site prediction are presented, as well as overall results for locating the 3' end of genes. In addition, future improvements to the algorithm are discussed, such as validation through proteomics data and improvement of start site prediction via signal peptide information.

Genetic Mapping of Genes Controlling Slow-Rust Resistance and Major Gene Resistance in Sugar Pine

Kathie Jermstad[1]* (kjermstad@fs.fed.us) and David Neale[2]

[1]USDA Forest Service, Institute of Forest Genetics, Placerville, California, and
[2]Department of Plant Sciences, University of California, Davis, California

Pines belonging to the *Pinus* subgenus *Strobus* are susceptible to a fungal pathogen (*Cronartium ribicola*) that was introduced to Northern America in the early 1900s. Surprisingly, several white pines have been shown to possess innate resistance to the rust infection. Two forms of resistance have been observed: 1) polygenic resistance (also known as quantitative or partial resistance) which expresses a wide distribution in disease phenotypes, and 2) monogenic resistance (also known as major gene resistance – MGR) which segregates in progeny as a single dominant gene. Breeding programs for deployment of the MGR are currently in place for sugar pine (*Pinus lambertiana*) however, resistance controlled by a suite of genes is likely to be more durable. Using sugar pine as a representative species, we are employing several strategies to develop genomic resources for the five-needled pines. Single nucleotide polymorphisms (SNPs) identified in 1200 loblolly pine *(P. taeda)* EST sequences from will be used for nucleotide diversity studies and comparative genomics within the Pinaceae (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber =0638502) including sugar pine. An EST resource is currently being developed for sugar pine through a grant recently awarded by the Joint Genome Initiative Community Sequencing Program (JGI-CSP_LOI_783956). Both of these EST resources will provide markers for genetic mapping in sugar pine. The mapping of quantitative trait loci (QTL) for partial resistance is underway in a large full-sib family (n>1000). In addition to mapping QTL for partial resistance, we are also pursuing a positional cloning strategy to isolate the DNA sequence (allele) conferring MGR. RAPD markers flanking *Cr*1 on the genetic map (Harkins et al. 1997) have recently been converted to Sequence Characterized Amplified Regions (SCARs) to facilitate in map-based cloning of *Cr*1. These markers could also provide an important diagnostic tool for studying range-wide population substructure and managing forest health.

Exploring Functions of Novel Proteins from an Acid Mine Drainage Biofilm

Yongqin Jiao* (yqchinster@gmail.com), Korin Wheeler, Steven Singer, Adam Zemla, Nathan VerBerkmoes, Robert Hettich, Daniela Goltsman, Jill Banfield, and Michael Thelen

Lawrence Livermore National Laboratory, Livermore, California

As metagenomic and proteomic studies expand, the number of genes and proteins of unknown function continually increases. However, high-throughput methods in analyzing the functions of these novel proteins are limited. Here we present a systematic approach to address this, using a model microbial community system from the Richmond Mine at Iron Mountain (Redding, CA)[1], that combines computational and experimental methods to predict function of several hundred novel proteins.

Protein sequences from a subset of 421 hypothetical genes from Leptospirillum group II, the dominant species in the microbial community, were analyzed using a structural modeling system (AS2TS)[2]. This resulted so far in the assignment of structural predictions for 360 proteins (85%). Experimentally, protein classes (e.g., hydrolases, oxidoreductases, phosphatases and amylases) and families (metalloproteins, tetratricopeptide [TPR] repeats) that are highly represented in the community are subject to enrichment from biofilm extracts via chromatography, or heterologous expression in *E. coli*. These samples are then tested for enzymatic activities. Moreover, protein complex analysis has been initiated using both a "tagless" approach and a bacterial two-hybrid screen.

[1]Ram et al, 2005, Science 308:1915-20, "Community Proteomics of a Natural Microbial Biofilm"

[2]Zemla et al, 2005, Nucleic Acids Res 33 (Web Server issue):W111-5, "AS2TS system for protein structure modeling and analysis"

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

An Annotation 'Minipipe' to Rapidly Assess Genomic Assemblies of 454 Pyrosequencing Reads

Alan Kuo[1]* (akuo@lbl.gov), Darren Platt,[1] Paul Richardson,[2] and Igor Grigoriev[1]

[1]Informatics and [2]Genomic Technologies, DOE Joint Genome Institute (JGI), Walnut Creek, California

The JGI Annotation Pipeline is routinely used to predict genes in assemblies of Sanger-sequenced eukaryotic genomes. In this study we instead use annotation to assess competing assemblies of a single genome sequenced using alternative technologies. These technologies promise to vastly lower the cost and thus expand the output of nucleotide sequencing compared to the traditional Sanger method. However, the new technologies also bring new challenges such as short reads and new kinds of sequence errors. At JGI we are exploring ways to incorporate 454 pyrosequencing into the standard JGI genome workflow. One way to exploit the respective strengths of 454 and Sanger sequence may be to derive 'hybrid' assemblies from both. To rapidly and consistently assess hybrid, 454-only (20X), and Sanger-only (4X) assemblies of the genome of the ubiquitous plant

pathogen *Phytophthora capsici*, we developed a short version of our standard JGI Annotation Pipeline. This 'minipipe' predicts genes, and than examines 1) internal stop codons and 2) truncated homologs of genes of Sanger-only (8X) *Phytophthora sp.* Both metrics serve as surrogates for frameshifts resulting from sequencing errors or polymorphism. Our results suggest that while the rate of 454-related frameshifting is high, the hybrid assembly approach holds promise as a low-cost way of estimating the number of genes in a genome. We conclude that annotation is a quick and simple way to assess the quality of novel assemblies.

Whole-Genome Gene Expression and Gene Family Analyses of *Phycomyces blakesleeanus*, A Model Photoresponsive Zygomycete

Alan Kuo* (akuo@lbl.gov), Asaf Salamov, Alex Atkins, Luis Corrochano, and Igor Grigoriev

Lawrence Berkeley National Laboratory, Berkeley, California

The photoresponsive zygomycete *Phycomyces blakesleeanus* is an experimentally tractable model system for elucidating the signalling pathways underlying photoregulation. Our recent sequencing, assembly, and annotation of the *P. blakesleeanus* genome allows us to complement classical molecular biological studies with genome-wide analyses of gene expression. We sequenced cDNAs from mycelia grown with and without light. The resulting 23410 'light' ESTs and 24437 'dark' ESTs were aligned with the genome and used to identify which of the 14792 genes could be described as transcriptionally active under either condition. We tagged 1538 genes (10%) as potentially light-specific, 1558 genes (11%) as potentially dark-specific, and 2431 genes (16%) as transcribed under both light and dark. In addition to its value as a model organism, *P. blakesleeanus* is one of only 2 zygomycetes with a sequenced genome, providing an opportunity to discover genes that are specific to or missing from zygomycetes, and gene families that are expanded or contracted in zygomycetes relative to the much better-sampled ascomycetes and basidiomycetes. We clustered *P. blakesleeanus* proteins with those of 13 other fungi and so far find 2 zygomycete-specific families of Zn-finger proteins, and a zygomycete-specific expansion of protein kinases. We also confirm the existence of a *Phycomyces*-specific family of F-box domain proteins.

Fungal Diversity in Hydrothermal Ecosystems and Predictions of Functions by a Metagenomic Analysis

Thomas Le Calvez* (thomas.lecalvez@univ-rennes1.fr) and Philippe Vandenkoornhuyse

University of Rennes 1, Rennes, France

Fungi are known as terrestrial (micro)organisms which have evolved and diversified in land. From molecular clock estimates we hypothesized their emergence and diversification in oceans. However, until now almost nothing is known about fungi in marine ecosystem. A current theory of emergence of life on earth suggests that the deep marine hydrothermal environment might be the cradle of life. Thus, we have tested and analyzed the diversity of fungi in this particular ecosystem. From ciPCR we have highlighted the presence of fungi in a variety of samples from different area around the world and the diversity was investigated by analyzing small subunit ribosomal RNA gene amplicons from total DNA extracts and from a culture collection that was successfully established. Our results

revealed an unsuspected diversity of species. We found a new branch of *Chytridiomycota* forming one of most ancient evolutionary lineage of fungi, in agreement with a our working hypothesis, the possibility of emergence and diversification of fungi in oceans conversely to the current dogma. The majority of the species found were new, even at higher taxonomic levels.

From this diversity study, a second appealing question was addressed, the ecological functions of these fungi in the hydrothermal ecosystem. In this aim, one particular sample was chosen on the basis of both fungal diversity and frequency by doing qPCR assays, on which a metagenomic analysis was performed by pyrosequencing (GS flx 454 Life Sciences, ROCHE).

Individual reads (450 Mb in total) were assembled *de novo* into 110 998 contigs, with an approximate coverage of 4x. A BLASTX search was performed against all contigs, with an e-value cutoff of 10^{-3}; all top hits (77677) were collected and classified according to their taxonomy. By this way, it was possible to establish a first taxonomic view of all living organisms' distribution in our sample. In a second time, we extracted the fungal contigs (335 non-redundant coding sequences), and re-analyzed them, performing BLASTn, BLASTX, and ORF finder to refine their characterization. Annotated Contigs were then examined in depth using a combination of KEGG, STRING and PFAM databases to predict their function and locate each protein coding gene in a metabolic networks. Given that nothing is known about fungal functions and metabolism in this ecosystem, we focused on the elucidation of metabolic pathways, such as energetic and carbohydrates metabolisms, which allowed us to establish hypotheses on fungal functions in this ecosystem. We found that they are able to make allelopathy by producing metabolites of the penicillin and streptomycine families. We also demonstrate their ability to produce most if not all aminoacids, allowing us to reject a pathogenic lifestyle. Until now, and surprisingly, we have no evidence or gene signature of heterotrophy for these fungi.

Analyses are in progress to resolve fungal metabolic processes by focusing on the Unidentified Open Reading Frames. We will also test the hypothesis of Horizontal Gene Transfers, to possibly highlight if they are primary or secondary colonizers of the hydrothermal ecosystem.

Quantitative Assessment of Phenotype and Pathway Conservation over Evolutionary Distance

Sara Light* (light6@llnl.gov.), Tomer Altman, and Patrik D'haeseleer

Computation Directorate, Lawrence Livermore National Laboratory, Livermore, California

16S rRNA similarity is often used to characterize environmental samples, under the implicit assumption that the role of an organism in the environment, i.e. its phenotype, is similar to that of its closest known relative. However, the utility of 16S as an approximation of phenotype may be questioned, as even strains from the same species can exhibit substantial phenotypic diversity. Here, we have critically assessed the relationship between evolutionary distance and phenotype, as well as pathway content, across the prokaryotes.

The 16S rRNA sequence alignments for all fully sequenced prokaryotic genomes were extracted from the GreenGenes database and the evolutionary distances between the 488 organisms were calculated. As an alternative measure of evolutionary distance we also used distances between organisms derived from a phylogenetic tree based on a selection of

31 marker genes. Phenotype data was collected from the NCBI, Genomes OnLine Database and TIGR's Comprehensive Microbial Resource, combined with a number of smaller datasets, curated, and eventually consolidated into 80 well defined and biologically relevant phenotypes. The average phenotypic similarity between pairs of organisms at different evolutionary distances shows significant conservation down to the class level, but with a large variance. However, some lineages show greater conservation of phenotype with evolutionary distance, and likewise, some phenotypes are much better conserved than others. Hence, although we cannot assume that closely related organisms will have similar phenotypes in the general case, it should be feasible to make predictions for some phenotypes and some lineages. Similarly, we have applied the method on 778 pathways covering 343 organisms, collected from BioCyc. As with the phenotypes, metabolic profiles show the highest degree of similarity between organisms of the same species or genus, and there is a moderate negative correlation between metabolic similarity and evolutionary distance. There is a much higher average similarity of metabolic profiles, presumably because some pathways are found in almost all genomes. Some lineages, such as Cyanobacteria, Chlamydiae and Archaea, have a significantly higher metabolic similarity within the lineage compared to other phyla. Likewise, some metabolic pathways are conserved across larger evolutionary distances.

Genome Dynamics Across Four New Pyrobaculum Species and Novel Non-Coding RNA Features

Todd Lowe* (lowe@soe.ucsc.edu), David Bernick, Patricia Chan, Aaron Cozen, and Matt Weirauch

Department of Biomolecular Engineering, University of California, Santa Cruz, California

Members of the crenarchaeal genus *Pyrobaculum* are widespread and abundant in neutral pH geothermal environments, and represent a unique clade among the Archaea because its members respire a large variety oxidants and have diverse metabolisms. The only member of this hyperthermophilic clade previously sequenced, *Pyrobaculum aerophilum*, contains a large proportion of genes that remain unverified and functionally obscure. Operon structures are unusually short relative to all other free-living archaea. Furthermore, extreme incidence of atypical tRNA introns and a “missing” universal RNase P RNA places *Pyrobaculum* in a class by itself.

Five species (*Pyrobaculum calidifontis*, *Pyrobaculum islandicum*, *Pyrobaculum arsenaticum*, *Thermoproteus [Pyrobaculum] neutrophilus*, *Caldivirga maquilingensis*) were selected and prepared by members of the *Pyrobaculum* Consortium for sequencing by the U.S. DOE Joint Genome Institute as part of the Community Sequencing Program. All five genomes are essentially complete and now publicly available from Genbank and the UCSC Archaeal Genome Browser (archaea.ucsc.edu). In addition to automated protein gene prediction provided by Oak Ridge National Labs, our lab has created a six-way full-genome alignment of the new genomes and *P. aerophilum*, and used a variety of comparative methods to observe gene gain/loss patterns in a study of genome dynamics, operon conservation, and genes unique but essential for this genus.

All genomes sequenced range in size between 1.76-2.12 million base pairs, with the two smallest, *P. islandicum* and *T. neutrophilus*, being most closely related and both from Iceland. The largest new genome, *P. arsentaticum*, is most similar in size and gene content to *P.aerophilum* -- both species were isolated in Italy. *P. calidifontis*, isolated in the Philippines and the only fully aerobic species sequenced, had a genome size falling in the

middle (2.0Mbp). All species share a common core of 1369 genes, 119 of which appear to be *Pyrobaculum*-specific, and may help define some of the unique characteristics of this group.

Full-genome dot-plots show significant shuffling of chromosomes, even between the most closely-related species. Although operons are short in all *Pyrobaculum* species, syntenic "neighborhoods" tend to remain adjacent to each other and on the same strand, even though they usually do not appear to be polycistronic. As such, we expect that conserved gene neighborhoods will have useful predictive value for determining gene function.

Furthermore, we found evidence for many novel DNA insertions relative to other genomes, presumably due to uncharacterized viruses. A majority of these insertions occurred near tRNA or C/D box sRNA genes. Transfer RNA genes were found to contain many dozens of "noncanonical" introns not conserved in other species, due to an apparently active insertion process not understood in mechanism or biological purpose. Upon close inspection of conserved intergenic regions among all five species, we found dozens of potentially novel RNA genes, one of which appears to be the "missing" RNase P gene. It was not detected before because one of two major functional domains has been lost.

In summary, *Pyrobaculum* sequencing has yielded the most detailed comparative genomics resource available among all Crenarchaea and hyperthermophiles, and is providing many leads for new *Pyrobaculum* research.

Chimera Free Insert Libraries for Next Generation Sequencing

David Mead* (dmead@lucigen.com), Rebecca Hochstein, Keynttisha Jefferson, Spencer Hermanson, and Ronald Godiska

Lucigen Corporation, Middleton, Wisconsin

The production of clone free libraries for DNA sequencing depends to a great extent on the quality of the random libraries produced. It is important to minimize chimeric inserts to facilitate accurate sequence assembly. Single-insert clones are usually ensured by ligating asymmetric linkers to the insert DNA. Subsequently, the excess linkers must be completely removed from the insert DNA before proceeding to the next step. We have developed a novel method to produce single inserts based on a new GC ligation technology and a unique DNA end blocking chemistry developed at Lucigen. GC ligation is analogous to TA ligations only a single 3'-C overhang is added to the adapter, which is compatible with the single dideoxy 3'-G overhang added to blunt ended DNA using PyroPhage DNA polymerase. The unique combination of a C tailed vector and ddG tailed insert blocks the ligation of multiple fragments. This protocol is robust and is significantly faster than TA ligations. The level of chimerism is ~ 1% in the libraries.

Linear *E. coli* Vector for Cloning Large PCR Products and Genomic DNA

David A. Mead[1]* (dmead@lucigen.com), Nikolai Ravin,[2] Sarah Vande Zande,[1] Rebecca Hochstein,[1] Karen Usdin,[3] Keynttisha Jefferson,[1] Spencer Hermanson,[1] and Ronald Godiska[1]

[1]Lucigen Corporation, Middleton, Wisconsin; [2]Centre BioEngineering RAS, Moscow, Russia; and [3]NIH, Bethesda, Maryland

A linear "pJAZZ" cloning vector for transformation of *E. coli* has been developed from the coliphage N15. This vector is used much like any plasmid but it provides exceptional ability to clone otherwise unclonable DNAs of any size up to ~30 kb. Modified versions of the vector allow for "GC cloning", analogous to TA cloning, but capable of cloning PCR fragments of up to 30 kb. The ends of the linear molecule are free to rotate during replication, so the vector and inserted DNA fragments are not subject to torsional stress caused by supercoiling. In addition, transcriptional terminators flank the cloning site, preventing transcriptional interference between the vector and insert. Regions of potential secondary structure appear to be very stable in this relaxed, non-transcribed state. For example, up to 300 copies (~ 1 kb) of the GCC repeat of the Fragile X locus were stable in the pJAZZ vector, but could not be maintained in a circular vector. AT-rich fragments of >20 kb and other highly repetitive DNAs also have been easily and reproducibly cloned and maintained through numerous rounds of sub-culturing. These regions were unstable in supercoiled plasmids, yielding deletions of both the circular vector and insert. The low bias of the pJAZZ vector greatly simplifies genomic sequence assembly due to the minimal number of gaps. In addition, GC cloning of PCR fragments using the pJAZZ vector is very useful for closing gaps in genomic libraries. The pJAZZ linear vector ensures unprecedented stability in maintaining large inserts in *E. coli*.

Analysis of Stress-Induced Changes in the Metabolism of *Yersinia pestis*

Ali Navid* (navid1@llnl.gov) and Eivind Almaas (almaas@llnl.gov)

Biosciences and Biotechnology Division; Chemistry, Materials, Earth, and Life Sciences Directorate; Lawrence Livermore National Laboratory, Livermore, California

The gram-negative bacterium *Yersinia pestis* is the aetiological agent of bubonic plague, a zoonotic infection that occurs through the bite of a flea. It has long been known that *Y. pestis* has different metabolic needs upon transition from the flea gut environment (26 °C) to that of a mammalian host (37 °C). To investigate the changes in genome-level metabolic activity and performance of *Y. pestis* when under duress, we used available genomic, biochemical and physiological data to develop a constraint-based flux balance model of metabolism in the avirulent 91001 strain (biovar Mediaevalis) of this organism. Utilizing sets of whole-genome DNA microarray expression data, we examined the system level changes that occur when *Y. pestis* acclimatizes to temperature shifts, or in response to anti-microbial agents (Chloramphenicol and Streptomycin). Our results point to fundamental changes in the oxidative metabolism of sugars and the use of amino acids, in particular those of arginine, serine and leucine.

This project was conducted under the auspices of United States Department of Energy at Lawrence Livermore National Laboratory (contract # W-7405-ENG-48), and was funded

by Laboratory Directed Research Development program (06-ERD-061). LLNL-ABS-401520.

Genetic Mapping and Physical Isolation of Telomeric and Subtelomeric Sequences in the Genome of the Basidiomycete *Pleurotus ostreatus*

Gumer Pérez, **Antonio G. Pisabarro*** (gpisabarro@unavarra.es), and Lucía Ramírez

Genetics and Microbiology Research Group, Department of Agrarian Production, Public University of Navarre, Pamplona, Spain

Pleurotus ostreatus (*Pleurotaceae*, *Agaricales*) is a model white-rot lignin-degrading basidiomycete whose importance is growing because of its ability to degrade selectively lignin in lignocellulosic substrates. This property makes *P. ostreatus* an attractive candidate for the development of bioethanol production processes, and of strategies for bioremediation of organic pollutants. The JGI is carrying out the whole *P. ostreatus* genome sequencing. The availability of this sequence will complement the information derived from the sequencing of the other white-rot basidiomycete (*Phanerochaete chrysosporium*) carried out by the JGI some years ago.

In order to determine the chromosomes borders, we have cloned, mapped and characterized telomeric and subtelomeric regions in the fungal strain being sequenced. The genome of *P. ostreatus* contains 11 chromosomes ranging in size from 1.4 to 4.7 Mbp, and there are prominent length polymorphisms in some of them. The telomeric sequence in this fungus is TTAGGG and we have determined that the length of the telomeres ranges from 150 to 1500 bp (25 to 250 repetitions of the basic unit). We have mapped 18 out of the 22 telomeres by RFLP using a $(TTAGGG)_{132}$ probe and stringent hybridization conditions. This approach has revealed cases in which that several telomeric regions map to the same chromosome, and others in which telomeric-hybridizing bands map to internal chromosome regions. Using a different strategy (SSP-PCR) we have cloned 38 telomere-containing fragments. Ten out of them consisted in monotone telomeric sequences up to a length of 55 repeats of the basic unit. The remaining 28 sequences corresponded to telomeric plus subtelomeric regions mapping to different chromosome ends and internal chromosome regions. One of them corresponded to an additional telomere that had not been mapped by RFPL, whose subtelomeric region was present in protoclon PC15 but missing in protoclon PC9. Three of these clones corresponded to telomeric-like sequences mapping to internal sites on the corresponding chromosomes (two of them mapping to chromosome III, and one to the highly polymorphic chromosome VI. Besides that, there are some cases in which the subtelomeric sequences are conserved in different chromosomes.

In summary, 19 out of the 22 expected telomeres have been genetically mapped. The study of the subtelomeric regions reveals a complex pattern of genome structure that can contribute to explain, at least partially, the chromosome length polymorphisms observed in this fungus.

Whole Genome Sequencing of the Two Spotted Spider Mite *Tetranychus urticae*: Novel Model for Plant-Herbivore Interactions

Cherise Poo,[1] Johannes Mathieu,[2] Richard Clark,[2] Marcus Schmid,[2] **Miodrag Grbic**[1]* (mgrbic@uwo.ca), and Vojislava Grbic[1]

[1]Department of Biology, University of Western Ontario, London, Ontario, Canada and [2]Max Planck Institute for Developmental Biology, Tuebingen, Germany

Application of chemical pesticides in agriculture represents one of the major costs of agricultural production and is a key source of environmental pollution, destruction of wildlife and introduction of carcinogens into the food chain.

Our current gap in knowledge about pest genetics, genomics and plant-pest interactions is a major obstacle for the development of alternative pest control strategies. To circumvent these problems we are developing new pest control methods by taking advantage of a novel pest genomic resource the whole-genome sequence of a major agricultural pest, the Two Spotted Spider Mite *Tetranychus urticae*.

Our goals are to:

1. Annotate the genome of the *T. urticae* and develop a spider mite whole genome expression microarray.
2. Analyze natural variation of plant resistance to spider mites using high-throughput genomic technologies.
3. Perform pest transcriptome profiling to characterize the consequences of feeding on resistant and susceptible plants.
4. Create pest-resistant transgenic plants targeting various pest genes.
5. Test the efficiency of the transgenic plants on pests and non-target organisms.
6. Analyze the pest and host plants in agro-ecosystem.

Our long-term goal is to develop environmentally sound pest control strategies that reduce environmental pollution and energy consumption in agriculture.

Key words: Genomics, pest control, plant biotechnology, pest resistance, biotic stress, bioinformatics, systems biology

Expansion of Protein Families and Gene Fate of Paralogs in the Ras and Protein Kinase Families of the Symbiotic Fungus *Laccaria bicolor*

Balaji Rajashekar,[1] Annegret Kohler,[2] Tomas Johansson,[1] Francis Martin,[2] Anders Tunlid,[1] and **Dag Ahrén**[1]* (dag.ahren@mbioekol.lu.se)

[1]Department of Microbial Ecology, Lund University, Lund, Sweden, and [2]UMR1136, INRA-Nancy Université, Interactions Arbres/Microorganismes, INRA-Nancy, Champenoux, France

Gene duplications and loss are major mechanisms generating evolutionary novelties and pruning specialized functions. Here, we have studied the role of duplication events for the evolution of ectomycorrhizae in *Laccaria bicolor*. Gene duplicates and gene families in the genomes of the *L. bicolor*, the saprophytes *Coprinopsis cinerea* and *Phanerochaete chrysosporium*, the human pathogen *Cryptococcus neoformans* and the plant pathogen

Ustilago maydis were analyzed. Among these basidiomycetes, *L. bicolor* contained the highest number of, mostly young, gene duplicates. The differences in gene duplicates had a pronounced effect on the number and size of multigene families. In total, 7352 protein families were identified in the five basidiomycete genomes. *L. bicolor* contained the largest number of lineage-specific (1077) and expanded (1064) protein families. A large fraction (29%) of the young gene duplicates of *L. bicolor* were found within the 55 largest gene families, having more than 25 members. Phylogenetic analyses of two such families, protein kinases and small GTPases, showed that *L. bicolor* contained clusters of paralogs that have arisen though duplication events in the *Laccaria* lineage. The gene fate after duplications was determined in the two families. Analysis of protein motif structures indicated that in the majority of the cases the ancestral motif structures had been retained in the paralogs. However, a few cases of changes in motif structures were observed that are indicative of neofunctionalization. Divergence in function of paralogs was also suggested by analysis of their expression levels in different tissues of *L. bicolor*.

Archaeal Virus Community Genomics of Yellowstone's High Temperature Environments

Frank Roberto[1]* (Francisco.Roberto@inl.gov), Alice Ortmann,[2] Mary Bateson,[2] Jennifer Fulton,[2] Josh Spuhler,[2] Trevor Douglas,[2] and Mark Young[2]

[1]Idaho National Laboratory, Idaho Falls, Idaho, and [2]Montana State University, Bozeman, Montana

We are testing the feasibility of describing the majority of viral diversity and virus community structure within an environment using a virus metacommunity sequencing approach. All metacommunity sequencing projects to date have encountered difficulties in assembling complete genomes due to the high complexity, relatively large genomes, and presence of DNA repetitive elements present in the environmental DNA samples. We have proposed to overcome these limitations by examining an archaeal virus environmental sample obtained from high temperature (80-90°C) acidic (pH 2-4) hot springs found in Yellowstone National Park (YNP). Initial analysis of one hot spring, CHAS, indicates that the viral community is of low complexity (estimated <10 different viruses types), with small genomes (<50kb), with little to no repetitive sequence. The value of the sequence information that will be generated by JGI will be maximized by linkage to a related microbial community metagenomics project (The Yellowstone Metagenome Project.[1]) A preliminary study of CHAS has provided support that this method will be successful. Analysis of the host-enriched metacommunity has revealed high rates of assembly of reads, suggesting a low complexity environment. These hosts appear to be dominated by Archaea. The virus-enriched metacommunity is more complex, with several large contigs showing signatures of known viruses, while a large proportion of the sequences show no similarity to known organisms. These unknown sequences represent a pool of novel genetic material to be mined for new viruses. Evidence of large, potentially conjugative plasmids has also emerged from early sequence information. The scientific value of these studies will be to greatly expand our understanding of archaeal viruses, their ecology (and influence on microbial host ecology) in extreme environments, and the role of virus migration on viral community structure.

[1]http://www.jgi.doe.gov/sequencing/why/CSP2008/yellowstone.html

Degradation of Polybrominated Diphenyl Ethers by Two PCB-Degrading Bacteria

Kristin R. Robrock[1]* (Robrock@berkeley.edu), Mehmet Coelhan,[3] David Sedlak,[1] and Lisa Alvarez-Cohen[1,2]

[1]Department of Civil and Environmental Engineering, University of California, Berkeley, California; [2]Ecology Department, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; and [3]Center of Food and Life Sciences, Technical University of Munich, Germany

Polybrominated Diphenyl Ethers (PBDEs) are commonly used flame retardants that have recently become a cause of concern because of their endocrine disruption ability and the high concentrations found in humans and in the environment. To date, very little is known about the potential for aerobic microorganisms to degrade these compounds. Here we investigate the ability of two polychlorinated biphenyl (PCB) degrading species, *Rhodococcus jostii* RHA1 and *Burkholderia xenovorans* LB400, to degrade a variety of PBDE congeners. The studies congeners three hexa-, three penta-, two tetra-, two tri-, two di-, and one mono-BDE, all added at 17ng/ml to the cultures. Both species are capable of degrading PBDE with five or fewer bromines within three days. The mono- and di-BDEs were completely degraded within three days, whereas only 20% of the penta-BDEs were degraded. RHA1 completely breaks down the PBDE molecule to release bromide. LB400 does not completely break down PBDEs, but instead forms a partially oxidized intermediate. The growth substrate greatly influences the ability of RHA1 to degrade PBDEs. Optimal degradation is seen when RHA1 is grown on biphenyl whereas almost no degradation is seen when grown on benzoate. The biphenyl (bph) and the two copies of the ethylbenzene (etb and edb) gene are involved with PCB degradation. It is likely that these genes are also involved with PDBE degradation as degradation patterns by growth substrate match their expression patterns.

Comparative Genomic Analysis of Transcriptional Regulatory Networks in *Shewanella* Species and Other γ-Proteobacteria

Dmitry A. Rodionov[1,2]* (rodionov@burnham.org), Dmitry Ravcheev,[2] Pavel Novichkov,[3] Elena Stavrovskaya,[2] Michael Cipriano,[4] Inna Dubchak,[4] Andrei Osterman,[1] and Mikhail Gelfand[2]

[1]Burnham Institute for Medical Research, La Jolla, California; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; [3]National Center for Biotechnology Information, Bethesda, Maryland; and [4]Lawrence Berkeley National Laboratory, Berkeley, California

Integrative comparative genomics approaches were used to infer transcriptional regulatory networks (TRNs) in 13 *Shewanella* species and a set of other γ-proteobacteria with sequenced genomes. To accomplish this goal, we combined the identification of transcription factors (TFs), TF-binding sites (TFBSs) and cross-genome comparison of regulons with the analysis of the genomic and functional context inferred by metabolic reconstruction. The reconstructed TRNs for the key pathways involved in central metabolism, production of energy and biomass, metal ion homeostasis and stress response provide a framework for the interpretation of gene expression data. This analysis also helps to improve functional annotations and identify previously uncharacterized genes in

metabolic pathways. Finally, we attempted to reconstruct possible evolutionary scenarios of these TRNs.

Using comparative genomics approach we identified candidate TFBSs for near eighty TFs from *Shewanella* group. For thirty described regulons, the TF was conserved between *E. coli* and *Shewanella.* These include global regulators (Crp, Fnr, ArcA, Fur, LexA) and specialized regulators of the metabolism of nitrogen (NarP, NsrR, DNR, NorR, NtrC), amino acids (ArgR, MetJ, TrpR, TyrR, HutC, IlvY, MetR), fatty acids (FadR, FabR), carbohydrates (SdaR, PdhR, HexR, GntR), and cofactors (BirA, IscR, ModE). Another fifty regulons described in *Shewanella* spp. are operated by TF that do not have orthologs in *E. coli*. In particularly, we characterized novel regulons that control gene involved in the degradation of branch chain amino acids (named LiuR), degradation of fatty acids (PsrA), and catabolism of various sugars (e.g. NagR, ScrR, AraR, and BglR tentatively implicated in the control of utilization of N-acetylglucosamine, sucrose, arabinose and β-glucosides, respectively).

Although some diversity of the predicted regulons is observed within the collection of *Shewanella* spp., the most striking difference in the overall regulatory strategy is revealed by comparison with *E. coli* and other γ-proteobacteria. Multiple interesting trends in diversification and adaptive evolution of TRNs between lineages were detected including regulon "shrinking", "expansion", "mergers", and "split-ups", as well as multiple cases of using nonorthologous regulators to control equivalent pathways or orthologous regulators to control distinct pathways.

Within the *Shewanella* lineage, the two major diversification strategies are: constrained ("all or none"), when the regulon is either present or absent in its entirety with tightly conserved regulation of all genes (e.g. for local regulons), and permissive ("loose"), when most genes of a regulon are conserved between genomes, whereas the conservation of respective regulatory sites is much weaker and sometimes not mandatory (e.g. for global regulons). Many aspects of metabolic regulation in *Shewanella* species are substantially different from TRN models that were largely derived from studies in *E. coli*. Among the most notable are the differences in TRNs for the central carbohydrate pathways. In enterobacteria the central carbon metabolism is controlled by catabolic regulators FruR and Crp, whereas *Shewanella* species use two other TFs, HexR and PdhR, for this control. The content and functional role of the Crp regulon is significantly different in these two lineages: the catabolism of carbohydrates and amino acids in enterobacteria, and the anaerobic respiration in *Shewanella* species.

Thauera sp. MZ1T: Preliminary Analysis of the Draft Genome Sequence

J. Sanseverino[1,2]* (jsansev@utk.edu), K. Jiang,[1,2] Y. Wang,[1] M.S. Allen,[1] D. Close,[1,4] K. Cusick,[1,2] J.M. DeBruyn,[1,3] A.C. Layton,[1,2] L. Poorvin,[1] and G.S. Sayler[1,2,3]

[1]Center for Environmental Biotechnology, [2]The Department of Microbiology, [3]The Department of Ecology and Evolutionary Biology, [4]Genome Science & Technology Program, The University of Tennessee, Knoxville, Tennessee

Thauera sp. strain MZ1T is a floc-forming bacterium isolated from the wastewater treatment plant of a major industrial chemical manufacturer. It is related to the genus *Azoarcus* and *Zoogloea*, another prominent community member of activated sludge. In previous research, MZ1T was identified as a significant component of clusters that resulted in poor sludge dewaterability. In pure culture, MZ1T produced copious quantities of EPS

from relatively simple short chain fatty acids. The draft genome was sequenced by the U.S. DOE Joint Genome Institute. The genome contains 4,518,936 base pairs in 94 contigs with a 68.3 percent GC content. There are 4,092 candidate protein-encoding gene models. The draft annotation of MZ1T indicated the complete glycolytic pathway and citric acid cycle is present as well as three key enzymes for assimilation of acetate: acetate-CoA ligase and acetate kinase - phosphate acetyl transferase.

Three extracellular polysaccharide (EPS) gene clusters have been identified. Two of the clusters are loosely arranged. The third cluster has a size of 20.67 kb and encodes 14 genes which include most of the genes necessary for EPS production and transport. Genes responsible for polysaccharide biosynthesis, polymerization, and export as well as chain length determinant have been identified. Currently, no regulatory elements have been identified. Most genes directing EPS biosynthesis (encoding glycosyl transferase, UDP-N-acetylglucosamine 2-epimerase, UDP-glucose/GDP-mannose dehydrogenase) have high homology (52 to 81% identity) to corresponding genes in various EPS producing bacteria such as *Pseudomonas aeruginosa.* Two genes involved in EPS chain length determination and export also have high homology (46 to 47% identity) to related genes in *Nitrosomonas europaea* ATCC 19718. The gene responsible for EPS polymerization has a 28% identity with the EPS polymerase gene in *Bacillus licheniformis* ATCC 14580.

Unlike closely related *Thauera* spp., MZ1T does not appear to have genes for anaerobic toluene or phenol degradation; however, genes for both anaerobic and aerobic benzoate degradation are present. The anaerobic degradation pathway proceeds via benzoyl CoA, with a benzoyl CoA degradation operon (*bcrCBAD*) very similar (93-98% identity) to that identified in *Thauera aromatica*. Genes that mediate the aerobic degradation of benzoate, phenol, and salicylaldehyde (all proceeding via catechol), along with intra- and extradiol ring cleavage dioxgyenases (similar to protocatechuate 3,4-dioxygenase and protocatechuate 4,5-dioxygenase) are present in MZ1T. Genes that mediate selenate reduction are also absent from this strain. Several transposase genes (including Tn3, and TNIS4) as well as prophage fragments have been tentatively identified in the genome.

QTL Mapping of Ligninolytic Activities in *Pleurotus ostreatus*

Santos F. Santoyo[1]* (santoyo.38899@e.unavarra.es), M.C.G. Terrón,[2] A.E. González,[2] L. Ramírez,[1] and A.G. Pisabarro[1]

[1]Genetics and Microbiology Research Group, Department of Agrarian Production, Public University of Navarre, Pamplona, Spain, and [2]Centro de Investigaciones Biológicas CIB-CSIC, Madrid, Spain

Pleurotus ostreatus is a model lignin-degrading basidiomycete. The lignin degrading strategy of this fungus includes phenol oxidases (Pox) and Mn-oxidizing peroxidases (Mnp/VP).

The purpose of this work is to compare the genetic linkage map positions of (1) the major QTLs for biomass production in solid and liquid cultures, and (2) of the QTLs controlling ligninolytic enzymatic activities with those of the corresponding structural genes. All these maps are based on the *P. ostreatus* linkage maps previously produced by our group (Larraya *et al.*, 2000, AEM **66**: 5290-5300). The QTLs analysis was made using in a population of 80 monokaryons, 274 molecular markers, and two open source programs (QTLCartographer V2.5 and MapQTL 5; Basten *et al.*, 1994, Zmap-a QTL cartographer. In 5th World Congress on Genetics Applied to Livestock Production: Computing

Strategies and Software. Guelph, Ontario, Canada, pp 65-66; Basten *et al.*, 2004, QTL Cartographer. Version 1.17. Department of Statistics. North Carolina State University).

The results obtained revealed a major QTL for biomass production mapping to chromosome IX, two major QTLs for MnP/VP activity (chromosomes VI and XI), and a major QTL for the Pox activity. The QTLs for biomass production in liquid culture mapped to different positions than those of solid cultures indicating the control by different genes in different environments; and some of the QTLs controlling enzymatic activities mapped to positions different than those of the structural genes suggesting that these regions might contain regulators of the corresponding enzymatic activities.

Currently, we are working on identifying cofactors (epistasis) for each character in order to make more robust analysis (CIM; composite interval mapping, MIM; multiple interval mapping). This part is being doing with a two-dimensional genome scan with a two-QTL model. These methods allow mapping increase the significance of each QTL, any time you identify genes that interact with them.

Advances in Thermophilic Phage DNA Polymerases and Their Implications for Emerging and Conventional DNA Detection and Analysis Methods

Thomas Schoenfeld, Vinay Dhodda, Innokenti Toulokhonov, and David Mead* (dmead@lucigen.com)

Lucigen Corporation, Middleton, Wisconsin

DNA polymerases are key components of most DNA amplification, sequencing and analysis platforms. As methods are developed and optimized for speed, throughput, accuracy and reliability, the need for improved enzymes is increasingly evident. Using the high throughput sequencing capacity of the JGI to discover novel enzymes as a starting point for engineering and directed evolution is a critical first step in addressing these needs. 37 Mb of sequence from thermophilic viral metagenomic libraries was screened resulting in the discovery of hundreds of thermostable viral DNA polymerases, as well as their respective accessory proteins. Ten *pol* genes were expressed to produce thermostable DNA polymerase activity. One of these allows unusually high-fidelity, high-efficiency PCR amplification of otherwise refractory sequences. Its thermostability and inherent stand-displacement activity allows isothermal synthesis at elevated temperatures, which results in greater specificity and lower background than is possible using conventional polymerases. Its inherent reverse transcriptase activity allows single-tube, single-enzyme RT PCR detection of mRNA transcripts. This enzyme also allows for the analysis of otherwise refractory sequences that compromise Sanger sequencing reactions and chain-termination SNP detection using existing enzymes. Improvements of other DNA polymerases and their impact on various DNA detection and analysis platforms will also be discussed.

Assembly of Viral Metagenomes from Yellowstone Hot Springs

Thomas Schoenfeld,[1] Melodee Patterson,[1] Paul M. Richardson,[2] K. Eric Wommack,[3] Mark Young,[4] and **David Mead**[1]* (dmead@lucigen.com)

[1]Lucigen Corporation, Middleton, Wisconsin; [2]Department of Energy Joint Genome Institute, Walnut Creek, California; [3]Department of Plant and Soil Sciences, University of Delaware, Newark, Delaware; and [4]Plant Sciences and Plant Pathology, Montana State University, Bozeman, Montana

Thermophilic viruses were first reported decades ago; however, knowledge of their diversity, biology and ecological impact is limited. Previous research on thermophilic viruses has focused on cultivated strains. This study examined metagenomic profiles of viruses directly isolated from 74° to 93°C mildly alkaline hot springs. Viral abundance ranged from 10^5 to 10^6 ml^{-1}. Using a new method for constructing libraries from picogram amounts of DNA, nearly 30 Mb of viral DNA sequence from two hot springs was determined at the JGI. Approximately 25% of the viral sequences share regions of significant similarity with the other hot spring. Although most sequences were unrelated to known genomes, hundreds of BLASTx similarities provide insights into viral lifestyles in this environment. In contrast to previous viral metagenomic studies, sequences were assembled at 50% identity, creating composite contigs as large as 35 kb that show the inherent heterogeneity in the populations. One 16.5 kb composite contig encodes 26 apparent virus-associated genes including three clones that express functional DNA polymerases. Lowering assembly identity from 95 to 50% reduced the number of different viral types from 1400 to 300. The 50% assembly included one contig of high similarity and perfect synteny to nine genes from *Pyrobacculum spherical virus* (PSV), a cultured thermophilic crenarchaeal virus. In fact, nearly all the genes of the 28 kb genome of PSV have apparent homologs in the metagenomes. Similarities to thermoacidophilic viruses isolated on other continents were limited to specific open reading frames but were equally strong. Metagenomics provides a powerful tool to study the diversity of viruses in these extreme environments.

A Randomized Approach to Evaluating Gene Characteristics and Overlap in the Green Lineage Marine Algae *Micromonas pusilla*

Melinda Simmons* (msimmons@mbari.org) and Alex Worden

Monterey Bay Aquarium Research Institute, Moss Landing, California

A significant fraction of terrestrial and oceanic carbon dioxide uptake is performed by green lineage organisms, including higher plants and green algae. The Prasinophyceae, a marine green algal taxa, are thought to have some similarities to the ancestral green flagellate {Lewis, 2004 #634}. Genome sequences, of the marine prasinophyte *Micromonas pusilla,* allow us to trace the genomic composition of the "proto-prasinophyte", and are reshaping how we think about the dynamics of these marine primary producers in the natural environment, as well as aspects of developmental and evolutionary biology. *M. pusilla* has a remarkably broad distribution, ranging from tropical to arctic oceans {Johnson, 1982 #329}{Raven, 1986 #643}. We sequenced the complete genomes of two *Micromonas* strains, one of which was isolated from the equatorial Pacific (RCC299) and the other (CCMP1545) from the North Atlantic. RCC299 has a 20,984,428 bp genome, which has been assembled telomere to telomere in 17 chromosomes. We sought to establish rigorous statistics on characteristics including gene size, exon number,

and gene overlap, a significant feature in this genome. To this end, the RCC299 genome was divided into 131 contiguous fragments (160,188 bp each) and then subsampled using a randomizing algorithm. Ten fragments were then selected for subsequent manual analysis of EST supported models. The analysis was also performed on one 160,188 bp non-randomly selected fragment located within a low %GC portion of Chromosome 1. This analysis allowed a more precise estimation of gene characteristics. While UTR and gene overlap has been noted in *Ostreococcus* (Derelle et al. 2006, Palenik et al. 2007), no statistics on this phenomena have been reported for the prasinophytes, and only a few studies have addressed it in other organisms. *Micromonas* strain RCC299 contains a higher percent of overlapping genes than reported to date for other eukaryotes.

The JGI Contribution to Whole Microbial Genome Sequencing

David Sims* (dsims@lanl.gov), Cliff Han, David Bruce, Tom Brettin, Chris Detter, and Cheryl Kuske

Los Alamos National Laboratory, Los Alamos, New Mexico

The DOE Joint Genome Institute (JGI) was created in 1997 to unite the expertise and resources in genome mapping, DNA sequencing, technology development, and information sciences pioneered at the DOE genome centers at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL).

Since that time, there has been an exponential increase in the number of whole microbial genome sequences that have been made publicly available through the National Center for Biotechnology Information (NCBI) as a result of the work of the JGI. There has been a similar increase in the proportion of the whole genomes that have been submitted by the JGI as compared to all other centers. To date, approximately 200 whole microbial genomes are available at www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1 that were completed by the JGI. This accounts for approximately 1/3 of all of the genomes, and twice the input of the next finishing center.

The JGI continues to strive to be the world leader in its contribution to this resource of basic microbiological knowledge, and to pave the way to next generation genome sequencing techniques and technologies.

Environmental Proteogenomics and Biochemistry Reveal a Possible Fe(II)-Dependent Electron Transfer Pathway for *Leptospirillum* Group II

Steven W. Singer[1]* (singer2@llnl.gov), Christopher Jeans,[1] Clara S. Chan,[2] Nathan C. VerBerkmoes,[3] Mona H. Hwang,[1] Paul F. Abraham,[3] Daniela Goltsman,[2] Paul Wilmes,[2] Manesh Shah,[3] Jillian F. Banfield,[2] Robert L. Hettich,[3] and Michael P. Thelen[1]

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

Interactions of microbes with metals underpin a variety of biogeochemical cycles. These cycles often involve redox-active proteins that mediate electron transfer reactions between extracellular metals and intracellular respiratory complexes. The availability of

proteogenomic data from natural samples has enabled electron transfer pathways to be inferred without isolation and culturing of target organisms. Acidophilic biofilms collected at the Richmond Mine in Iron Mountain, CA readily oxidize the abundant, soluble Fe(II) in mine water. The dominant species in many of these biofilms, *Leptospirillum* group II, is difficult to maintain in culture and little information is available about its Fe(II) oxidation pathway. Proteogenomics of these biofilms has identified two highly abundant cytochromes with novel amino acid sequences encoded by the genome of *Lepto*II. These cytochromes were purified directly from biofilm samples and shown to have unusual covalently-bound heme groups. Mass spectrometric analysis of the purified proteins revealed that amino acid variation and post-translational modification of these abundant cytochromes were dependent on the ecological state of the biofilms from which they were purified. Proteogenomics also identified several *Lepto*II *c*-type cytochromes present at lower levels in acidic protein extracts of the biofilm. These *c*-type cytochromes were enriched by column chromatography and visualized by SDS-PAGE heme stains. Amino acid variants of some of these *c*-type cytochromes were observed in ecologically distinct biofilms, and these variants are being expressed as heterologous proteins in *E. coli* to determine if the variation alters the biochemical properties of the cytochromes. From these data, an Fe(II)-dependent electron transfer pathway for *Lepto*II is proposed. This proposed pathway reflects the dynamic interaction of *Lepto*II with the extreme environment in which it thrives.

Microbial and Genetic Analysis of a Microbial Community Actively Decaying Poplar Biomass

Safiyh Taghavi,[1] Susannah Green Tringe,[2] Tanja Woyke,[2] Shi-You Ding,[3] Michael Himmel,[3] and **Daniel van der Lelie**[1]* (vdlelied@bnl.gov)

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

As primary decomposers, microbial communities have evolved as competitors and collaborators in deconstructing biomass. To understand and exploit these complex microbial communities and their dynamics for the conversion of recalcitrant plant biomass to useful bioenergy feedstocks, a highly integrated research initiative is required. A prerequisite is to have insight into the composition and metabolic potential of this lignocellulosic biomass degrading community. To achieve this goal several complementary strategies are used to study the microbial community actively decaying poplar woodchips:

Analysis of total community composition: After the isolation of metagenome DNA, 16S and 18S rRNA genes were PCR amplified, shotgun cloned and sequenced. The distribution of the species showed that members of the order *Clostridiales*, many of which are closely related to uncultivatable bacteria, comprise 85% of this community. The presence of the majority of members of this order in the community is expected because of the mesophylic, anaerobic conditions characteristic for the sample. *Saccharomycetes* composed the major group among the Eukaryotes.

Isolation and characterization of cultivable microorganisms: the high number of uncultivable microorganisms in this consortium was confirmed by cultivation studies. None of the cultivable bacteria represented the dominant members of the community as determined via 16S RDNA sequencing. Isolated strains are presently being screened for their glycosylhydrolase activity.

Metagenome sequencing: in order to obtain a thorough understanding of the diversity, structure, functional interdependence, and metabolic capabilities of this community. This approach is providing unprecedented insights in the diversity of glycosylhydrolases present in plant biomass decomposing microbial communities.

Synergistic Interactions Between Poplar and Endophytic Bacteria to Improve Plant Establishment and Sustainable Feedstock Production on Marginal Soils

Safiyh Taghavi, **Sebastien Monchy*** (smonchy@bnl.gov), and Daniel van der Lelie (vdlelied@bnl.gov)

Brookhaven National Laboratory, Biology Department, Upton, New York

Producing biomass that is tailored toward energy production, but that does not negatively impact food supply is one of the critical social-economical issues of the proposed U.S. biofuel program. Poplar, considered as the model tree species for bioenergy feedstock production, live in close association with symbiotic microorganisms.

By screening approximately 100 poplar's endophytes, we showed that specific bacteria had a beneficial effect on poplar biomass production (up to a 30% increase). Among them, four strains were sequenced (*Enterobacter* sp. 638, *Stenotrophomonas maltophilia* R551-3, *Pseudomonas putida* W619 and *Serratia proteamaculans* 568) in order to better understand the complex interactions between endophytes and poplar. Genome annotation, analysis of metabolic properties, and genome comparisons with closely related non-endophytic bacteria resulted in the identification of several unique pathways, by which endophytic bacteria can promote plant growth and health. Those pathways include the production of phytohormones such as indole-3-acetic acid, acetoin and diacetyl. The expert genome annotation of *Enterobacter* sp. 638 reveals the presence of many genes, important for bacteria/plant interactions, encoding for transporters involved in the uptake of sugar, amino acids and iron (siderophore), and proteins required for plant colonization such as adhesin, agglutinin, and pili. These genes are often located on genomic islands. *Enterobacter* sp. 638 also contains a plasmid of 158 kb that carries four clusters of genes possibly involved in plant colonization. The expression of these genes during colonization is being studied using Q-PCR and microarrays.

In addition, the comparison between the recently sequenced genome of *Populus trichocarpa* and its endophytic partners will result in a better understanding of the synergistic interactions between plant and bacteria. Those data can be exploited to improve plant establishment, growth and health in order to sustain bioenergy feedstock production on marginal, non-agricultural land.

The Complete Genome Sequence of *Thermosinus carboxydivorans* str. Nor1: Linking Hydrogenogenic Carboxydotrophy to Metal Reduction

Stephen Techtmann[1]* (techtman@umbi.umd.edu), Albert Colman,[2] Linda Meincke,[3] Jonathan Eisen,[4] and Frank T. Robb[1]

[1]University of Maryland Biotechnology Institute, Center of Marine Biotechnology, Baltimore, Maryland; [2]University of Chicago, Department of Geophysical Science, Chicago, Illinois; [3]Department of Energy Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico; and [4]University of California, Davis, California

Many microbes have the ability to utilize carbon monoxide (CO) as their sole carbon and energy source. A sub-class of these organisms, known as hydrogenogens, couples the oxidation of CO with reduction of water to form CO_2 and H_2. Thermophilic microbes have dominated the recently isolated members of this physiology. *Thermosinus carboxydivorans* is a strictly anaerobic extreme thermophile that was isolated from Norris Basin in Yellowstone National Park. It has the unusual ability to couple the oxidation of CO to the reduction of ferric iron or selenite. When the organism is oxidizing CO and reducing Fe(III) it continues to produce copious amounts of hydrogen. Unlike the other sequenced thermophilic hydrogenogen – *Carboxydothermus hydrogenoformans* – this bacterium is not an autotroph. This unique physiology has led us to sequence the genome of *T. carboxydivorans* to attempt to understand the physiology of hydrogenogenic carboxydotrophy and to better understand the coupling of CO oxidation with metal reduction.

The genome of *T. carboxydivorans* is 2.9 Mb, has a %GC content of 51% and is currently in closure with five contigs. *T. carboxydivorans'* genome contains three carbon monoxide dehydrogenases (CODHs), which are the enzyme complexes that oxidize CO to CO_2. Conspicuously absent from the genome is a CODH acetyl CoA synthase which enables the fixation of carbon from CO or CO_2. Its absence explains the fact that *T. carboxydivorans* cannot grow on CO autotrophically. The three CODHs are interesting when compared with the five found in *C. hydrogenoformans.* The catalytic subunit of the hydrogenase-linked CODH has 90% identity on the nucleotide level to the CODH I of *C. hydrogenoformans.* Genome-wide, sequence identity between *C. hydrogenoformans* and *T. carboxydivorans* is 82%. Another CODH is unlinked to other CO-related genes and is highly similar to the only CODH in *C. hydrogenoformans* that does not have a putative function. The third CODH is a minimal CODH composed of only 482 amino acids compared to the typical CODHs that have 612-722 amino acids. This minimal CODH is associated with one other CO-related gene, *cooF*. Additionally, the CO-responsive transcriptional regulator (CooA) is not linked to any CODH operon and is flanked by transposases and phage related genes. These findings lead us to hypothesize that *T. carboxydivorans* acquired the carbon monoxide related genes by way of horizontal gene transfer. These findings also support the conclusion that *T. carboxydivorans* is much more specialized for hydrogenogenic carboxydotrophy and does not utilize CO for multiple fates like *C. hydrogenoformans.*

Two more interesting findings are related to iron metabolism and sporulation. The genome contains three Fur-like regulators, two *feoAB* gene cassettes, two versions of an iron ABC transporter substrate-binding protein, and a gene encoding the intracellular iron storage protein, ferritin. This demonstrates genomically the ability of *T. carboxydivorans* to reduce iron. In terms of sporulation, the description paper reports that this organism is non-spore forming, but the genome contains all of the necessary genes for sprorulation initiation.

This indicates the previously unknown potential of *T. carboxydivorans* for endospore formation.

Analysis of Diversity in New *Brachypodium distachyon* Germplasm Resources

John Vogel[1]* (jvogel@pw.usda.gov), Hikmet Budak,[2] Metin Tuna,[3] Daniel Hayden,[1] and Michael Steinwand[1]

[1]USDA-ARS Western Regional Research Center, Albany, CA; [2]Sabanci University, Orhanli, Tuzla-Istanbul, Turkey; and [3]Namik Kemal University, Tekirdag, Turkey

Brachypodium distachyon (*Brachypodium*) is a small grass with all the attributes needed to be a modern model organism including simple growth requirements, fast generation time, small stature, small genome size and self-fertility. *Brachypodium* serves as a complementary model to Arabidopsis, in that it allows researchers to address questions about biological features unique to the grasses (e.g. cell wall composition). In the past few years a significant number of genomic resources have been developed to facilitate the utilization of *Brachypodium* as a model system. These include: facile *Agrobacterium*-mediated transformation protocols with efficiencies on par with rice, BAC libraries, a physical map, a genetic linkage map containing 200 PCR-based markers, and chemically (EMS) and fast neutron mutagenized populations. In addition, a high density genetic map containing ~1,000 SNP markers will be completed shortly and a project to scale up the generation of sequence indexed T-DNA mutants has been initiated. When combined with the whole genome sequence currently in production at JGI, the case for using *Brachypodium* for many applications becomes compelling. In contrast to other *Brachypodium* resources, freely available germplasm is currently very limited. To develop new germplasm resources we made >500 collections of *Brachypodium* seed from 100 locations across Turkey and are creating >500 inbred lines that will be freely distributed to the research community. *Brachypodium* has a natural range centered around the Mediterranean extending north into Europe and south into the Indian subcontinent. Within this region it occupies a variety of habitats including hot interior regions, cooler coastal areas and colder mountainous regions. Not surprisingly, traits relevant to biofuel crops such as seed size, flowering time, vernalization requirements and disease resistance vary considerably. In preliminary studies, we have examined the genetic and phenotypic variability of 68 diploid lines. This collection includes six inbred lines derived from USDA collections from Iraq and Turkey and 62 new inbred lines from 8 locations in Turkey. Turkey is a particularly rich source of *Brachypodium* diversity because it covers a variety of habitats including coastal regions, hot interior deserts and cold northern highlands. Phenotypically, this collection contained lines that required vernalization times ranging from 2 to 12 weeks, seed sizes ranging from 2.5 to 5.9 mg/seed, hairless seeds and hairy seeds. Further, we examined genetic diversity using 44 SSR markers and found significant genotypic variation both between and within populations. The molecular and phenotypic diversity observed in this collection indicates that *Brachypodium* will be a suitable model for applications that utilize natural diversity to understand gene function and regulation. Thus, *Brachypodium* is a excellent candidate for a whole genome resequencing project and our results could be used to select genotypes for resequencing. A summary of our characterization of this *Brachypodium* collection will be presented.

The University of Minnesota BioFuels Database

Lawrence P. Wackett[1]* (wacke003@umn.edu), Lynda Ellis,[2,3] Marc vonKeitz,[2] Carol Gross,[1,2] Naomi Kreamer,[1,2] and Pradeep Narayanashetty[2]

[1]Department of Biochemistry, Molecular Biology and Biophysics; [2]BioTechnology Institute; and [3]Laboratory Medicine and Pathology, University of Minnesota, St. Paul, Minnesota

Society is currently transitioning from petroleum-based to biomass-based chemicals and fuels. This transition is both exciting and chaotic. To help ease this transition, we have developed the University of Minnesota Biofuels Database (UM-BFD):

http://www.biofuelsdatabase.org

The UM-BFD is a freely available, internet resource. It builds on our more than 12 years experience in developing the Biocatalysis/Biodegradation Database, a resource describing novel microbial enzymes and metabolic pathways. The UM-BFD also describes enzymes and pathways, but contains other information relevant to fuel synthesis. The UM-BFD helps address the following questions: (1) What types of molecules are potential fuels? (2) What metabolic pathways can make fuel molecules? (3) How do chemical and biological synthetic routes compare? The core of UM-BFD information is organized around chemical compounds that are presently, or proposed to be, used as fuels. The fuels are alcohols, esters, ethers and hydrocarbons. Specific fuels within each category have pathway pages with a synthetic scheme for making them, chemically or biochemically.

The current biofuel market is dominated by alcohols (ethanol) and esters (biodiesel). However, biofuels with superior overall properties, for example hydrocarbons, may emerge as preferred choices. To expand the choices, and to develop ideal fuels from renewable resources, users will need information on relevant biosynthetic reactions. As new routes to novel fuel molecules are developed, society will need to consider economic and environmental implications. For example, users may investigate the likely environmental fate of any new fuel molecule by linking from the UM-BFD to the UM-Pathway Prediction System (PPS). The UM-PPS predicts how fuels, and other compounds that might contaminate the environment, are metabolized by microorganisms:

http://umbbd.msi.umn.edu/predict/

With these different functionalities, the UM-BFD is poised to become an important resource for those interested in biofuels research and development.

Discovery of Genes for Improved Cellulose and Cellulose-Extractability from Poplar Secondary Xylem

Jill L. Wegrzyn* (jlwegrzyn@ucdavis.edu), Jennifer M. Lee, Andrew J. Eckert, Charlyn Suarez, and David B. Neale

University of California, Davis, California

The DOE's "Breaking the Biological Barriers to Cellulosic Ethanol" report identifies poplar as one of the key feedstock species for cellulosic ethanol production in many regions of the country. The goal was to perform SNP discovery using high throughput DNA sequencing (Agencourt Biosciences) and SNP genotyping (Illumina) to associate genetic variation in genes involved in cellulose and lignin biosynthesis with phenotypic

variation in cellulose quantity, quality and extractability in a large clonal black cottonwood (*Populus trichocarpa*) genetic test plantation belonging to GreenWood Resources. A set of 40 genes known to be highly expressed and associated with the desired phenotypes were sequenced using a panel of 15 unrelated poplar clones. Genomic sequences of ~179,000 bp covering the entire protein-coding regions, including introns, and 1,000bp upstream and 300bp downstream were retrieved from JGI. 200 non-overlapping amplicons were selected to cover the length of the genes and were sequenced by Agencourt in both directions and submitted to an automated pipeline developed in-house for sequence alignment and SNP discovery. Utilizing Illumina's Golden Gate Assay, 456 poplar clones were genotyped for ~1536 SNPs. Wood samples were collected from 1,100 trees and came from the 456 poplar clones. High-throughput phenotyping has been performed with pyrolysis molecular beam mass spectrometry to analyze wood chemistry components such as lignin, cellulose, and hemicellulose on the cores collected. Association genetics analysis will be used to identify genes controlling cellulose quantity and quality phenotypic variation in poplar.

Single Cell Genome Reconstruction of an Uncultured, Proteorhodopsin-Containing Flavobacterium

Tanja Woyke* (TWoyke@lbl.gov), Alex Copeland, Gary Xie, Cliff Han, Jan-Fang Cheng, Hajnalka Kiss, 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 hinders the assembly of individual genomes from metagenomic shotgun libraries. Here we report the use of single cell genomics to access the genome of a proteorhodopsin-encoding flavobacteria from Gulf of Maine bacterioplankton. We use high throughput fluorescence-activated sorting of single cells, whole genome amplification via multiple displacement amplification, PCR-screening and subsequent shotgun sequencing of this single amplified genome (SAG), allowing the genomic analysis of its novel photometabolic system and the sequence comparison to environmental marine sequence data.

Random Shear BAC Libraries for Improved Genome Finishing

Cheng-Cang Wu, Rosa Ye, Becky Hochstein, Keynttisha Jefferson, Ronald Godiska, and **David A. Mead*** (dmead@lucigen.com)

Lucigen Corporation, Middleton, Wisconsin

Bacterial artificial chromosome (BAC) libraries and BAC-based physical maps are a vital resource for assembling large complex genomes. DNA libraries built with conventional vectors and methods are biased, resulting in numerous gaps in all of the physical and sequencing maps that have been produced to date. To overcome cloning bias and gaps due to partial restriction digestion, we have successfully developed techniques to construct unbiased, physically sheared BAC libraries with large inserts (>100 kb). Over a dozen random shear BAC libraries have been completed including important model species such as *Xenopus tropicalis* and *Medicago truncatula*. We will present gap closing data using the *Arabidopsis* genome. To further reduce cloning bias we have developed a unique

BAC/fosmid cloning system, termed “pSMART BAC v2.0”. This BAC/fosmid vector lacks an indicator gene and associated promoter, and has termination signals on both sides of the insert. The vector shows much higher stability of inserts containing AT-rich sequences, direct and inverted repeats, and other deleterious DNAs, thus making it possible to construct unbiased large-insert libraries. The CopyRight pSMART BAC v2.0 vector and the associated Replicator v2.0 competent cells feature inducible amplification of copy number, increasing yields to as many as 50 copies per cell. The amplification is more robust than alternative systems and permits easy isolation of BAC DNA for sequencing, subcloning, or restriction mapping. We are also implementing a vector barcode for multiplex library sequencing and to help track and sort libraries. The combination of vector improvements and random shearing library construction allows genome finishing to be done more efficiently, economically and completely.

Identification of Significant Temporal Motifs in Biological Networks

Sooyeon Yoon[1]* (yoon6@llnl.gov), Ruoming Jin,[2] and Eivind Almaas[1]

[1]Bioscience and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California, and [2]Department of Computer Science, Kent State University, Ohio

The recent interest in modeling biological systems as complex networks has generated many important insights. However, one shortcoming of the majority of such studies is that the network links and nodes are considered to be of equal importance and unchanging with time. For many biological systems, it is natural to represent the disparity in link and node importance with weights. The change in these weights with time can be among the most important factors to understand mechanisms governing pattern formation and system organization.

We have developed a novel set of efficient data-mining tools, trend motifs, to extract significant patters from weighted networks that are evolving with time. In this study, we generalize the definition of trend motifs, and apply it to several networks, in particular protein-interaction networks to elucidating how they organize temporally in response to stresses such as heat-shock and hypo-osmotic shock.

This project was conducted under the auspices of United States Department of Energy at Lawrence Livermore National Laboratory (contract # W-7405-ENG-48), and was funded by Laboratory Directed Research Development program (06-ERD-061).

UCRL number: LLNL-ABS-401334

Nanogram DNA Sample Preparation for Next Generation Sequencing

Tao Zhang* (tzhang3@lbl.gov), Matt Blow, Len Pennacchio, and Eddy Rubin

DOE Joint Genome Institute, Walnut Creek, California, and Genome Sciences Division, University of California, Lawrence Berkeley National Laboratory, Berkeley, California

The sample preparation for the next generation GS-FLX and Illumina GA sequencing typically requires microgram level of starting DNA for building sequencing library. However, certain low-DNA-content biological samples, such as ancient DNA and

chromatin immuno-precipitation (ChIP) enriched DNA, are unable to reach this benchmark and need pre-amplification before sequencing. We have developed an adaptor ligation mediated emulsion PCR amplification method to amplify nano-gram quantity of DNA. In this process, DNA molecules ligated with linkers are amplified clonally in water-in-oil droplets to minimize the bias caused by conventional PCR amplification. Because the linkers are identical to the ones used for library preparation in the next generation sequencing platforms, the amplified library is fully compatible to those sequencers. We consistently obtain over five hundred-fold amplification of the starting DNA materials from nanogram amount of ancient DNA extracts. Analysis of over one hundred megabases of sequences from the amplified 40,000 years old ancient wolf DNA products reveals that the redundancy in these amplified libraries is low and the amplification is unbiased. This method has also been applied to sequence nanogram quantity of ChIP-DNA for mapping histone modification in mammalian genome, and nanogram of microbial cDNA for transcriptom analysis.

Genomic and Expression Profiles of Switchgrass-Targeted Rumen Microbial Communities

Tao Zhang[1,2]* (tzhang3@lbl.gov), Susannah Tringe,[1,2] Matthias Hess,[1,2] Tanja Woyke,[1,2] Jennifer Heguy,[3] Ed DePeters,[3] and Eddy Rubin[1,2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Genome Sciences Division, University of California, Lawrence Berkeley National Laboratory, Berkeley, California; and [3]Department of Animal Science, University of California, Davis, California

Conversion of cellulosic biomass into fermentable sugars is the major bottleneck in bio-ethanol production. Currently this process relies on chemical and physical pretreatments that are expensive, energy consuming, and environmentally harmful. Yet in the gut environments of ruminants the degradation of raw plant feedstock materials via enzymatic reactions takes place rapidly and at moderate pH and temperature. We aim to identify enzymes produced by the rumen microbes responsible for degradation of plant cell wall polymers that can be co-opted for cellulosic biomass conversion. The fistulated cow offers easy, nondestructive access to the rumen, making it possible to inoculate bioenergy feedstocks in the rumen using synthetic bags. In this study, fragmented switchgrass leaf and straw fibers are placed in nylon bags and incubated in the rumen for 72 hours. Both DNA and RNA are extracted from microbial communities colonized on the switchgrass fibers, or free floating in the rumen fluid. 16S rRNA-based community profiles indicate that feedstuffs can influence the rumen microbial community, and cDNA transcript profiles reveal the expression of genes involved in fiber degradation, carbohydrate metabolism and other rumen biochemical functions.

Assessment of the ABI SOLiD High-Throughput Sequencing Technology

Zhiying Jean Zhao* (zyzhao@lbl.gov), **Kanwar Singh*** (ksingh@lbl.gov), Steven Wilson, Paul Winward, Paul Richardson, Len A. Pennacchio, and Feng Chen (fchen@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

Short read DNA sequencing technologies are poised to dramatically decrease sequencing cost and are expected to have suitable roles in tasks such as variation detection and error correction in genome assemblies. At the U.S. DOE Joint Genome Institute, we have been actively assesses the features of several such platforms. Here we described our ongoing experience with the ABI SOLiD system. This technology utilizes a ligation based sequencing scheme and is anticipated to generate several gigabases of mappable data per run. Our preliminary results indicate that the whole process is quite robust and is capable of generating 1.5 Gb per run. We performed error and genome coverage analysis from several microbes as well as two Arabidopsis ecotypes. Our findings will be reported. Ongoing challenges remain in the areas of molecular workflow and operation time but over time these obstacles are likely to be overcome.

Attendees

Current as of March 11, 2008

Andrea Aerts
DOE Joint Genome Institute
alaerts@lbl.gov

Dag Ahren
Lund University
dag.ahren@mbioekol.lu.se

Andrew Ellis Allen
J. Craig Venter Institute
aallen@jcvi.org

Eivind Almaas
Lawrence Livermore National Lab
almaas2@llnl.gov

Amy Anderton
USDA-ARS GGD WRRC
aanderton@pw.usda.gov

John M. Archibald
Dalhousie University
jmarchib@dal.ca

Christopher E. Bagwell
Savannah River National Lab
christopher.bagwell@srnl.doe.gov

Scott E. Baker
Pacific Northwest National Lab
scott.baker@pnl.gov

Jill Banfield
Univ. of California, Berkeley
jbanfield@berkeley.edu

Kerrie Barry
DOE Joint Genome Institute
kwbarry@lbl.gov

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

Khela Baskett
DOE Joint Genome Institute
kbweiler@lbl.gov

Diane Bauer
DOE Joint Genome Institute
dmbauer@lbl.gov

Jason K. Baumohl
DOE Joint Genome Institute
jkbaumohl@lbl.gov

Bonnie K. Baxter
Westminster College
bbaxter@westminstercollege.edu

Gry Mine Berg
Stanford University
mineberg@stanford.edu

Stephanie M. Bernard
LBNL
smbernard@lbl.gov

David Bernick
Univ. of California, Santa Cruz
dbernick@soe.ucsc.edu

Jeffrey Boore
Genome Project Solutions
Univ. of California, Berkeley
jlboore@genomeprojectsolutions.com

Jennifer N. Bragg
USDA, ARS, WRRC
jbragg@pw.usda.gov

Susan Brawley
University of Maine
brawley@maine.edu

Terry Bristol
Institute for Science,
Engineering and Public Policy
bristol@isepp.org

Pamela Jane Bonner Brown
Indiana University
pjbonner@indiana.edu

Shane A. Brubaker
LS9, Inc.
brubaker@ls9.com

Yves V. Brun
Indiana University
ybrun@indiana.edu

Gregory Butler
Concordia University
gregb@cs.concordia.ca

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

Romy Chakraborty
LBNL
rchakraborty@lbl.gov

Srinivasa Rao Chaluvadi
University of Georgia
src@uga.edu

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

Yun-juan Chang
Oak Ridge National Lab
yjs@ornl.gov

Feng Chen
DOE Joint Genome Institute
fchen@lbl.gov

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

Mansi Chovatia
DOE Joint Genome Institute
mrchovatia@lbl.gov

Julianna Chow
DOE Joint Genome Institute
jchow@lbl.gov

Alicia N. Clum
DOE Joint Genome Institute
aclum@lbl.gov

Frank Collart
Argonne National Lab
fcollart@anl.gov

Eileen Dalin
DOE Joint Genome Institute
e_dalin@lbl.gov

Chris Daum
DOE Joint Genome Institute
daum1@llnl.gov

Jeffrey Dean
University of Georgia
jeffdean@uga.edu

Patrik D'haeseleer
Lawrence Livermore National Lab
patrikd@llnl.gov

Shi-You Ding
National Renewable Energy Lab
shi_you_ding@nrel.gov

Victor Dorsett
DOE Joint Genome Institute
vtdorsett@lbl.gov

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

Jennifer C. Drew
University of Florida
jdrew@ufl.edu

Aecio D'Silva
University of Arizona
ams@email.arizona.edu

Inna Dubchak
JGI, LBNL
ildubchak@lbl.gov

Attendees

Erin Dunwell
DOE Joint Genome Institute
dunwell2@llnl.gov

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

Jonathan A. Eisen
JGI, UC Davis
jaeisen@ucdavis.edu

Helene Feil
Univ. of California, Berkeley
bhfeil@nature.berkeley.edu

Marsha Fenner
DOE Joint Genome Institute
mwfenner@lbl.gov

Klaus Fiebig
Ontario Genomics Institute
kfiebig@ontariogenomics.ca

Susan I. Fuerstenberg
Genome Project Solutions
sifuerst@genomeprojectsolutions.com

Craig Furman
DOE Joint Genome Institute
cfurman@lbl.gov

Elisabeth Gantt
University of Maryland
egantt@umd.edu

Audrey P. Gasch
Univ. of Wisconsin, Madison
Great Lakes Bioenergy Res. Center
agasch@wisc.edu

Cheol-Min Ghim
Lawrence Livermore National Lab
cmghim@llnl.gov

Carol S. Giometti
Argonne National Laboratory
csgiometti@anl.gov

John Glass
J. Craig Venter Institute
jglass@jcvi.org

Robert Glass
Lawrence Livermore National Lab
glass3@llnl.gov

Jacob Y. Golder
DOE Joint Genome Institute
jgolder@lbl.gov

Eugene Goltsman
DOE Joint Genome Institute
egoltsman@lbl.gov

Dario Grattapaglia
EMBRAPA Brazilian
Federal Research Corporation
dario@cenargen.embrapa.br

Miodrag Grbic
Univ. of Western Ontario
mgrbic@uwo.ca

Susan K. Gregurick
U.S. Department of Energy
susan.gregurick@science.doe.gov

Annette Greiner
DOE Joint Genome Institute
amgreiner@lbl.gov

Igor Grigoriev
DOE Joint Genome Institute
ivgrigoriev@lbl.gov

Jane Grimwood
Stanford Human Genome Center
jane@shgc.stanford.edu

Arthur Robert Grossman
Carnegie Institution
arthurg@stanford.edu

Yong Qiang Gu
USDA-ARS, WRRC
yong.gu@ars.usda.gov

Yalong Guo
Max Planck Institute
ya-long.guo@tuebingen.mpg.de

Fred Gvillo
fred_gvillo@msn.com

Ping Hu
LBNL
phu@lbl.gov

Christopher Alan Hack
DOE Joint Genome Institute
cahack@lbl.gov

Matthew Hamilton
DOE Joint Genome Institute
mghamilton@lbl.gov

James Han
DOE Joint Genome Institute
jkhan@lbl.gov

Shunsheng Han
Los Alamos National Lab
han_cliff@lanl.gov

Justin D. Hatch
Freezer tech
jhtatch@lbl.gov

Loren John Hauser
Oak Ridge National Lab
hauserlj@ornl.gov

Samuel P. Hazen
Univ. of California, San Diego
shazen@ucsd.edu

Terry C. Hazen
Lawrence Berkeley National Lab
tchazen@lbl.gov

Jianzhong He
National University of Singapore
jianzhong.he@gmail.com

Charles Douglas Hershberger
Codexis
hersh006@gmail.com

Matthias Hess
DOE Joint Genome Institute
mhess@lbl.gov

Mike Himmel
National Renewable Energy Lab
mike_himmel@nrel.gov

Isaac Y. Ho
DOE Joint Genome Institute
iyho@lbl.gov

Krassimira Hristova
Univ. of California, Davis
krhristova@ucdavis.edu

Matthew Hudson
University of Illinois
mhudson@uiuc.edu

Naxin Huo
USDA-ARS, WRRC
nhuo@pw.usda.gov

Karla Ikeda
DOE Joint Genome Institute
kmikeda@lbl.gov

William Inskeep
Montana State University
binskeep@montana.edu

Natalia Ivanova
DOE Joint Genome Institute
nnivanova@lbl.gov

Janet K. Jansson
Lawrence Berkeley National Lab
jrjansson@lbl.gov

Kathleen Jermstad
USDA-Forest Service, PSW
kjermstad@fs.fed.us

Yongqin Jiao
Lawrence Livermore National Lab
yqchinster@gmail.com

Magnus Karlsson
Forest Mycology & Pathology, SLU
magnus.karlsson@mykopat.slu.se

Matthew Kaser
Bell & Associates
mkaser@bell-iplaw.com

Lisa Kegg
DOE Joint Genome Institute
kegg2@llnl.gov

Martin Keller
Oak Ridge National Lab
kellerm@ornl.gov

Cheryl Kerfeld
JGI, UC Berkeley
ckerfeld@lbl.gov

Edwin Kim
DOE Joint Genome Institute
ekim@lbl.gov

Eunsoo Kim
Dalhousie University
eunsookim@dal.ca

James Kinney
DOE Joint Genome Institute
jkinney@lbl.gov

Michael Geoffrey Klein
DOE Joint Genome Institute
mgklein@lbl.gov

Frank Korzeniewski
DOE Joint Genome Institute
frkorzeniewski@lbl.gov

Anthony Kosky
DOE Joint Genome Institute
askosky@lbl.gov

Alan Kuo
DOE Joint Genome Institute
akuo@lbl.gov

John Kyndt
University of Arizona
jkyndt@email.arizona.edu

Peter Lammers
New Mexico State University
plammers@nmsu.edu

Miriam Land
Oak Ridge National Lab
landml@ornl.gov

Dorothy Lang
Lawrence Livermore National Lab
lang21@llnl.gov

Christina Lanzatella-Craig
USDA PWA WRRC
ccraig@pw.usda.gov

Alla Lapidus
DOE Joint Genome Institute
alapidus@lbl.gov

Gerard R. Lazo
USDA-ARS
lazo@pw.usda.gov

Thomas Le Calvez
University Rennes 1
thomas.lecalvez@univ-rennes1.fr

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

Susan Leschine
Univ. of Massachusetts Amherst
suel@microbio.umass.edu

Mingkun Li
DOE Joint Genome Institute
mli@lbl.gov

James C. Liao
Univ. of California, Los Angeles
liaoj@ucla.edu

Sara Kristina Light
Lawrence Livermore National Lab
light6@llnl.gov

Erika Lindquist
DOE Joint Genome Institute
ealindquist@lbl.gov

Stephen Long
University of Illinois
slong@uiuc.edu

Todd Lowe
Univ. of California, Santa Cruz
lowe@soe.ucsc.edu

Steve Lowry
DOE Joint Genome Institute
slowry@lbl.gov

Susan Lucas
DOE Joint Genome Institute
lucas11@llnl.gov

Athanasios Lykidis
DOE Joint Genome Institute
alykidis@lbl.gov

Chris Mackenzie
University of Texas
Health Science Center
ronald.c.mackenzie@uth.tmc.edu

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

Shaily Mahendra
Rice University
mahendras@rice.edu

Stephanie Malfatti
JGI, LLNL
malfatti3@llnl.gov

Lisa Margonelli
New America Foundation
margonelli@newamerica.net

Konstantinos Mavrommatis
DOE Joint Genome Institute
kmavrommatis@lbl.gov

Stephen Mayfield
The Scripps Research Institute
mayfield@scripps.edu

David Mead
Lucigen
dmead@lucigen.com

Juan C. Meza
Lawrence Berkeley National Lab
jcmeza@lbl.gov

David Mills
Univ. of California, Davis
damills@ucdavis.edu

Debra Mohnen
University of Georgia
dmohnen@ccrc.uga.edu

Sebastien Monchy
Brookhaven National Laboratory
smonchy@bnl.gov

Jenna Morgan
JGI, UC Davis
jlmorgan@lbl.gov

Ali Navid
Lawrence Livermore National Lab
navid1@llnl.gov

Simona Necula
DOE Joint Genome Institute
sfnecula@lbl.gov

Abby Ngau
DOE Joint Genome Institute
wcngau@lbl.gov

Matt Nolan
DOE Joint Genome Institute
mpnolan@lbl.gov

Pavel Novichkov
NCBI/NLM/NIH
psnov@ncbi.nlm.nih.gov

Donald Lee Nuss
University of Maryland
nuss@umbi.umd.edu

Howard Ochman
University of Arizona
hochman@email.arizona.edu

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

Robert Otillar
DOE Joint Genome Institute
rpotillar@lbl.gov

Krishnaveni Palaniappan
JGI, LBNL
kpalaniappan@lbl.gov

Bernhard Palsson
Univ. of California, San Diego
palsson@ucsd.edu

Attendees

Jasmyn Pangilinan
DOE Joint Genome Institute
jlpangilinan@lbl.gov

Georgios Pappas
EMBRAPA
gpappas@cenargen.embrapa.br

Andrew Paterson
Univ Georgia
paterson@uga.edu

Ashtamurthy Pawate
Sandia National Laboratories
apawat@sandia.gov

Yi Peng
DOE Joint Genome Institute
ypeng@lbl.gov

Ze Peng
DOE Joint Genome Institute
zpeng@lbl.gov

Len Pennacchio
DOE Joint Genome Institute
lapennacchio@lbl.gov

Rene Perrier
DOE Joint Genome Institute
raperrier@lbl.gov

Jennifer Pett-Ridge
Lawrence Livermore National Lab
jeffiner@nature.berkeley.edu

Antonio G. Pisabarro
University of Navarre
gpisabarro@unavarra.es

Samuel Pitluck
DOE Joint Genome Institute
s_pitluck@lbl.gov

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

Reno Pontarollo
Genome Prairie
fpagdonsolan@genomeprairie.ca

Simon Prochnik
DOE Joint Genome Institute
seprochnik@lbl.gov

Theodore K Raab
Stanford University
tkraab@stanford.edu

Preethi Ramaiya
Novozymes, Inc.
pira@novozymes.com

Gary Resnick
Los Alamos National Lab
tanyal@lanl.gov

Paul Richardson
DOE Joint Genome Institute
pmrichardson@lbl.gov

Monica Riley
Marine Biological Lab
mriley301@comcast.net

Frank Robb
UMBI
robb@umbi.umd.edu

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

David Robinson
DOE Joint Genome Institute
dsrobinson@lbl.gov

Kristin Regan Robrock
Univ. of California, Berkeley
robrock@berkeley.edu

Dmitry Rodionov
Burnham Institute
rodionov@burnham.org

Nina Rosenberg
Lawrence Livermore National Lab
rosenberg4@llnl.gov

Asaf A. Salamov
DOE Joint Genome Institute
aasalamov@lbl.gov

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

Manoj P. Samanta
Systemix Institute
manoj.samanta@systemix.org

John Sanseverino
University of Tennessee
jsansev@utk.edu

Francisco Santoyo
Genetics and Microbiology Res. Group
santoyo.38899@e.unavarra.es

Nori Satoh
Kyoto Univeristy
satoh@ascidian.zool.kyoto-u.ac.jp

Wendy Schackwitz
DOE Joint Genome Institute
wsschackwitz@lbl.gov

Jeremy Schmutz
Stanford Human Genome Center
jeremy@shgc.stanford.edu

Falk Schuetzenmeister
TU Dresden
postfalk@web.de

Igor Shabalov
DOE Joint Genome Institute
ishabalov@lbl.gov

Henry F. Shaw
Lawrence Livermore National Lab
shaw4@llnl.gov

Jay A. Shendure
University of Washington
shendure@u.washington.edu

Melinda Simmons
Monterey Bay Aquarium Res. Inst.
msimmons@mbari.org

David Sims
Los Alamos National Lab
dsims@lanl.gov

Steven W. Singer
Lawrence Livermore National Lab
singer2@llnl.gov

Kanwar Singh
DOE Joint Genome Institute
ksingh@lbl.gov

Mitchell L. Sogin
Marine Biological Laboratory
sogin@mbl.edu

Michael Steinwand
USDA ARS
msteinwand@pw.usda.gov

John William Stiller
East Carolina University
stillerj@ecu.edu

Hui Sun
DOE Joint Genome Institute
hsun@lbl.gov

Sirisha Sunkara
DOE Joint Genome Institute
ssunkara@lbl.gov

Aijazuddin Syed
DOE Joint Genome Institute
asyed@lbl.gov

Ernest Szeto
JGI, LBNL
eszeto@lbl.gov

John W. Taylor
Univ. of California, Berkeley
jtaylor@nature.berkeley.edu

Stephen Techtmann
Univ. of Maryland Biotech. Institute
techtman@umbi.umd.edu

Sarah A. Teter
Novozymes, Inc
sate@novozymes.com

Michael P. Thelen
Lawrence Livermore National Lab
mthelen@llnl.gov

Hope Tice
DOE Joint Genome Institute
tice1@llnl.gov

Damon Tighe
DOE Joint Genome Institute
tighe2@llnl.gov

Tamas Torok
Lawrence Berkeley National Lab
ttorok@lbl.gov

Susannah Green 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

Timothy Tschaplinski
Oak Ridge National Lab
t2t@ornl.gov

Hank Tu
DOE Joint Genome Institute
hctu@lbl.gov

Jerry Tuskan
ORNL, JGI
gtk@ornl.gov

Jun Urano
Gevo, Inc.
jurano@gevo.com

Anna Ustaszewska
DOE Joint Genome Institute
ustaszewska1@llnl.gov

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

Meric Velasco
DOE Joint Genome Institute
mvelasco@lbl.gob

John Vogel
USDA-ARS
jvogel@pw.usda.gov

Humphrey W
USDA-ARS
hwanjugi@yahoo.com

Lawrence P. Wackett
University of Minnesota
wacke003@umn.edu

Virginia Walbot
Department of Biology
walbot@stanford.edu

Mei Wang
DOE Joint Genome Institute
mwang@olbl.gov

Jill Wegrzyn
University of California, Davis
jlwegrzyn@ucdavis.edu

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

Gary Xie
Los Alamos National Lab
xie@lanl.gov

Yifeng Yin
yifeng_yin@yahoo.com

Wendy Thompson Yoder
Novozymes, Inc
wty@novozymes.com

Sooyeon Yoon
Lawrence Livermore National Lab
yoon6@llnl.gov

Frank M. You
USDA-ARS-WRRC-GGD
frankyou@pw.usda.gov

Curtis Robert Young
Harvard University
cyoung@oeb.harvard.edu

Michael Yu Zhang
DOE Joint Genome Institute
myzhang@lbl.gov

Tao Zhang
DOE Joint Genome Institute
tzhang3@lbl.gov

Zhiying Zhao
DOE Joint Genome Institute
zyzhao@lbl.gov

Kemin Zhou
DOE Joint Genome Institute
kzhou@lbl.gov

Gerben Zylstra
Rutgers University
zylstra@aesop.rutgers.edu

Author Index

Authors

Notes


JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE