Fifth Annual
DOE Joint Genome Institute
User Meeting

Sponsored By

U.S. Department of Energy
Office of Science

March 24–26, 2010

Walnut Creek Marriott

Walnut Creek, California

Contents

Speaker Presentations

Abstracts alphabetical by speaker

Solving Problems With Sequences

Rita Colwell (rcolwell@umiacs.umd.edu)

University of Maryland, College Park

Genome Insights Into Early Fungal Evolution and Global Population Diversity of the Amphibian Pathogen *Batrachochytrium dendroabatidis*

Christina Cuomo (cuomo@broadinstitute.org)

Broad Institute, Cambridge, Massachusetts

Batrachochytrium dendrobatidis (*Bd*) is a fungal pathogen of amphibians implicated as a primary causative agent of amphibian declines. The genome sequence of *Bd* was the first representative of the early diverging group of aquatic fungi known as chytrids. With the JGI, we have sequenced and assembled the genomes of two diploids strains: JEL423, isolated from a sick *Phylomedusa lemur* frog from Panama and JAM81, an isolate from Sierra Nevada, CA. By identifying polymorphisms between these two assemblies with survey sequence from 5 additional global isolates, we have characterized the genome-wide pattern of variation, and used conservation patterns between strains to predict the structure of an ancestral genome. Further, fixed polymorphisms in the otherwise homozygous regions can be used to estimate the divergence date of the sequenced strains. By comparing the predicted proteins of *Bd* to that of other fungi and eukaryotes, we identified gene families expanded in *Bd*, some with potential roles in pathogenesis. The recent sequence of two additional chytrid genomes allows more specific characterization of such gene families within chytrids, and better delineation of expansions in the lineage leading to *Bd*. We have also characterized a set of genes conserved only with non-fungal eukaryotes, some of which play a role in cilia or flagella in those species. Comparative analysis with the additional chytrid genomes will strengthen this basal vantage point for genomic comparisons across the fungi as well as with the sister animal clade and other eukaryotes.

The Promise and Challenge of Producing Biofuel Feedstocks: An Ecological Perspective

Evan Delucia (delucia@life.illinois.edu)

University of Illinois, Urbana-Champaign

From Fjords to Open Seas: Ecological Genomics of Expanding Oxygen Minimum Zones

Steven Hallam (shallam@interchange.ubc.ca)

Department of Microbiology and Immunology, University of British Columbia, Vancouver, Canada

Oxygen minimum zones (OMZs), also known as oceanic "dead zones", are widespread oceanographic features currently expanding due to global warming. Although inhospitable to metazoan life, OMZs support a thriving but cryptic microbiota whose combined metabolic activity is intimately connected to nutrient and trace gas cycling within the global ocean. Therefore, OMZ expansion and intensification represents an emerging ecological phenomenon with potentially harmful effects on ocean health and climate balance. In order to understand, respond to, or mitigate these transitions, studies monitoring and modeling dynamics and systems metabolism of OMZ microbiota in relation to physical and chemical oceanographic parameters are imperative. To this end we are using time-resolved metagenomic approaches to chart microbial community responses to changing levels of water column hypoxia in the Eastern Subtropical North Pacific Ocean (ESTNP). The ESTNP is one of the world's most extensive OMZs and provides an exceptional model system for long-term observation and process-oriented studies of OMZ phenotypes.

Genomics on the Half Shell: So, What Do Oysters Have to Do with Energy?

Dennis Hedgecock[1] (dhedge@email.usc.edu), Andrew Y. Gracey,[1] and Patrick M. Gaffney[2]

[1]Department of Biological Sciences, University of Southern California, Los Angeles and [2]School of Marine Science and Policy, University of Delaware, Lewes

For more than a decade, the Pacific oyster *Crassostrea gigas* has had the highest annual production of any aquatic organism on the planet, 4.2 million metric tons in 2007, according to the FAO. Like crops and other plants being considered as feedstocks for biofuels production, the Pacific oyster shows dramatic heterosis or hybrid vigor. Whereas the genetic and physiological causes of heterosis in major crops have remained mysteries for over 100 years, in the Pacific oyster, a large load of mutations and efficiency of protein metabolism and balanced gene expression are emerging as explanations for growth heterosis.

A JGI community sequencing project for the oyster was completed in 2009, with the delivery of high quality, Sanger paired-end sequencing reads for 72.5k cDNA clones, nearly a million 454 sequences from larval cDNA, and finished sequences from 58 BAC clones from the Pacific oyster (*Crassostrea gigas*) and two BACs from the eastern oyster (*C. virginica*).

The objectives of the EST sequencing were gene discovery and SNP identification. Sanger cDNA sequences were obtained mostly from mixed adult tissue libraries comprising tagged and pooled cDNA from two inbred lines. The average length of cDNA clones is 2.5 kb; the average length of EST sequences, 678 bp. Sequences have been clustered and aligned against a variety of databases, including a *C. gigas* EST database (http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas/index.html) and the genome of the limpet *Lottia gigantea* (http://genome.jgi-psf.org/Lotgi1/Lotgi1.home.html). Nearly 12,000 clones, representing unigenes with good BLAST hits, are being re-sequenced by Illumina

Genome Analyzer to provide a full-length cDNA resource. SNP candidates are currently being generated for analysis by Illumina Golden Gate Assay in mapping families and a panel of divergent stocks and closely related species.

The objective of the BAC sequencing was to assess genomic polymorphism. BAC clones were selected on the basis of positive hybridization to probes for eight genes of immunological interest. For each set of BACs, contigs have been assembled, representing one or two haplotypes. Alignment of allelic contigs reveals numerous macroindels and polymorphisms, the former often representing active transposable elements of various classes, the latter including microindels, micro- and minisatellites, and single nucleotide polymorphisms. The extensive polymorphism observed is comparable to that reported for the ascidian *Ciona*, and is causing difficulties in assembling nearly 100× genome coverage obtained by Illumina sequencing at the Beijing Genome Institute.

Competitiveness of Second Generation Biofuel Feedstocks: Role of Technology and Policy

Madhu Khanna (khanna1@illinois.edu)

University of Illinois, Urbana-Champaign

There is growing interest in using perennial grasses and crop residues for producing ligno-cellulosic ethanol. To engender a viable bio-based energy system, perennial grasses must compete successfully both as crops and as fuels. The competitiveness of various bioenergy crops and crop residues for biofuels as compared to corn ethanol in the U.S. is analyzed. Crop productivity models are used to simulate the yields of the bio-energy crops and life-cycle analysis is used to determine their potential to reduce greenhouse gas emissions relative to gasoline. The break-even prices needed to produce these feedstocks and the spatial variability in these prices across the U.S. is examined. We also examine the implications of alternative biofuel subsidy and carbon price policies and the role of technological innovation in improving the competitiveness of these feedstocks.

Title To Be Determined

Steve Knapp

University of Georgia, Athens

Phenotypes Phirst! – Tackling Complex Traits in the Field

T. Mitchell-Olds* (tmo1@duke.edu), K. Prasad, B.H. Song, C. Olson-Manning, C.R. Lee, A. Manzaneda, and J. Anderson

Institute for Genome Sciences and Policy, Department of Biology, Duke University, Durham, North Carolina

The expanding power and throughput of genomics is beginning to illuminate the genetic causes and evolutionary significance of complex trait variation. Building on these advances, plant biologists face the challenge of measuring, understanding, and predicting phenotypic variation in natural and agricultural environments. Towards this goal, we

discuss ecological genetics of a cloned QTL in natural plant populations, and then consider pathways forward in crop improvement.

By positional cloning we identified a QTL controlling insect resistance and plant defense in wild populations of *Boechera stricta*, a wild relative of *Arabidopsis*. In the original populations where these alleles evolved, we measured insect resistance and individual fitness. The causal polymorphism predicts trait variation in the wild, experiences strong ecological selection, and shows molecular signatures of non-neutral evolution. Taking an interdisciplinary approach, we discuss the evolutionary history of the causal polymorphism, its ecological effects, and the catalytic consequences of amino acid changes.

Tackling the Triple-Threat Genome of *Miscanthus x giganteus*

Stephen P. Moose (smoose@illinois.edu)

Department of Crop Sciences, University of Illinois, Urbana-Champaign

Miscanthus x giganteus (*Mxg*) is a highly productive perennial grass species that shows considerable promise as a dedicated bioenergy crop. *Mxg* is a member of the *Andropogoneae* tribe, which also includes the leading bioenergy crops maize, *Saccharum* (sugarcane), and *Sorghum*. Each of these crops performs the highly efficient C4 form of photosynthesis that also enhances water and nutrient use efficiency. *Mxg* offers additional advantages of continued C4 photosynthesis in chilling temperatures (<10^{o}C), recycling of nutrients from shoots to underground rhizomes in late-fall prior to harvest, and is a sterile triploid that favors greater biomass accumulation due to hybrid vigor and lack of seeds, which also reduces concerns about invasiveness. Our Feedstock Genomics Program within the Energy Biosciences Institute began in 2008 with the goal of developing the sequence resources that will enable genomics-directed improvement of *Mxg* and its close relatives as a bioenergy crop. From a starting point of sequences deposited in GenBank for one gene (a key enzyme in C4 photosynthesis) and the rDNA spacer, we have used 454 and Illumina deep sequencing technologies to generate a skim survey of the *Mxg* genome, to profile *Mxg* small RNAs, and to assemble millions of ESTs from *Mxg* and representatives of its progenitor species, *Miscanthus sinensis* and *Miscanthus sacchariflorus*. We will present an update on these efforts, plans to complete draft assemblies of the *Mxg* genomes, and insights into the interesting biology of *Miscanthus*.

Metabolic Noise, Vestigial Metabolites or the Raw Material of Ecological Adaptation? Opportunistic Enzymes, Catalytic Promiscuity and the Evolution of Chemodiversity in Nature

Joseph P. Noel (noel@salk.edu)

Howard Hughes Medical Institute and The Salk Institute for Biological Studies, La Jolla, California

We are mapping the adaptive molecular changes that have occurred in enzymes and metabolic pathways of specialized metabolism as these enzymes and enzyme networks emerged and subsequently evolved from their ancestral roots in primary metabolism billions of years ago. Specifically, we experimentally measured the first systematic quantitative characterization of a catalytic landscape underlying the evolution of sesquiterpene chemical diversity. Based on our previous discovery of a set of 9 naturally occurring amino acid substitutions that functionally inter-converted orthologous

sesquiterpene synthases from *Nicotiana tabaccum* and *Hyoscyamus muticus*, we created a library of all possible residue combinations (29 = 512) in the *N. tabaccum* parent. The product spectra of 418 active enzymes reveal a rugged landscape where several minimal combinations of the 9 mutations encode convergent solutions to the inter-conversions of parental activities. Quantitative comparisons indicate context dependence for mutational effects – epistasis – in product specificity and promiscuity. Furthermore, this sesquiterpene skeletal complexity in nature also originates from the enzyme-catalyzed ionization of (trans, trans)-farnesyl diphosphate (FPP) and subsequent cyclization along either 2,3-transoid or 2,3-cisoid farnesyl cation pathways. The aforementioned, tobacco 5-epi-aristolochene synthase (TEAS), a transoid synthase, produces cisoid products as a component of its minor product spectrum. To investigate the cryptic cisoid cyclization pathway in TEAS, we employed (cis, trans)-FPP as an alternative substrate. Strikingly, TEAS was catalytically robust in the enzymatic conversion of (cis, trans)-FPP to exclusively (≥ 99.5%) cisoid products. Further, crystallographic characterization of wild-type TEAS and a catalytically promiscuous mutant (M4 TEAS) with 2-fluoro analogues of both all trans FPP and (cis, trans)-FPP revealed binding modes consistent with pre-organization of the farnesyl chain. These results provide a structural glimpse into both cisoid and transoid cyclization pathways efficiently templated by a single enzyme active site, consistent with the recently elucidated stereochemistry of the cisoid products. What possible relevance does the cryptic cisoid cyclization pathway of TEAS have in the natural world? Although (cis, trans)-FPP has not been identified as a metabolite in tobacco or related Solanaceous plants, a (cis, trans)-farnesyl diphosphate synthase has been identified in Mycobacterium tuberculosis involved in bacterial cell wall synthesis suggesting the potential relevance of this compound in other biological systems. Moreover, while often observed, the biological significance of small amounts (3-14% of total product) of (cis, trans)-FPP formation by FPP synthases has been ignored to date. Is it possible then that TEAS possesses a "moonlighting" role in vivo by gathering up what we would normally consider biosynthetic 'waste' and recycling it into a bioactive product?

Patterns of Nitrogen Utilization in Deep-Sea Syntrophic Consortia

Victoria Orphan* (vorphan@gps.caltech.edu), Anne Dekas, and Abigail Green

Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena

Genes and Genomics for Improving Energy Crops

Roger Pennell

Ceres, Thousand Oaks, California

Dedicated energy crops such as switchgrass and miscanthus have only a short history of improvement by breeding. Yet they hold considerable promise as sustainable sources of feedstock for the bioenergy industries. To support a rapid program of improvement, genomics-based genetic studies are required. Genes of known function are also to be transformed and could be very valuable. Ceres has created large full-length plant cDNA libraries and annotated the genes. Individual such genes were selected and transformed into Arabidopsis, and the plants screened to find the genes that made improvements in biomass and other specific traits. A subset of these genes was then transformed into rice and screened in the field in China. These studies revealed genes that enhanced biomass and other traits in both species, and these are now being evaluated in switchgrass and other

energy crops in the USA. This program of genomics-based gene selection for energy crop improvement will be described.

Ocean Viral Metagenomics

Forest Rohwer

San Diego State University, San Diego, California

Three Non-Technical Challenges in the Development of Biomass-Based Energy

Steve Savage (ssavage@cirruspartners.com)

Cirrus Partners, Evergreen, Colorado

For a new, biomass-based energy industry to progress, it will not only have to overcome technical and economic hurdles, but also structural and socio-political challenges as well. One challenge will be the sufficient adoption of these crops by growers. The uneven adoption of continuous no-till practices around the world will be considered as an example of relevant barriers and drivers. A second issue will be the increasing "para-regulatory" environment of market access restrictions. These are being driven using "sustainability metrics" as a means to elicit consumer brand protectionism. A third challenge will be the increasing politicization of science. This development undermines the potential for biomass-based energy to benefit from the monetization of externalities that once appeared to have the potential to send logical economic signals driven by life-cycle-analysis.

Soybean

Gary Stacey

University of Missouri, Columbia

Fungal Decomposition of Lignocellulosic Substrates

Adrian Tsang (tsang@gene.concordia.ca)

Department of Biology and Centre for Structural and Functional Genomics, Concordia University, Montreal, Québec, Canada

Our project aims to enhance the understanding of the strategies used by fungi in the decomposition of lignocellulose, and to use the knowledge gained to devise enzyme cocktails for the efficient hydrolysis of agricultural and forestry residues. In this presentation, I will use *Myceliophthora thermophila* (*Sporotrichum thermophile*) and *Thielavia terrestris*, the genomes of these two thermophilic fungi have been sequenced recently by the Joint Genome Institute, as examples to illustrate the steps that we take to decipher the extracellular proteins used by fungi in breaking down lignocellulose. 1) All genes predicted to encode lignocellulolytic enzymes are manually curated and compare to their orthologues in other species. 2) The fungi are cultured using agricultural straws or thermo-mechanical pulp as nutrients. 3) The RNAs are extracted for analysis using the

RNA-Seq method of the Illumin/Solexa platform. 4) Extracellular proteins are identified by mass spectrometry. 5) All genes predicted to encode predicted lignocellulolytic enzymes and genes encoding unknown proteins identified in the exo-proteome are cloned and expressed in a heterologous host. The recombinant enzymes are characterized for basic biochemical properties. 6) The transcriptome and exo-proteome profiles are used to guide the development of enzyme cocktails. 7) The recombinant enzymes are tested for their utility in industrial applications. Results from these ongoing activities will be presented. Also discussed are results from two related projects: 1) the development of a database of manually curated fungal genes encoding characterized lignocelluloses-degrading enzymes, and 2) the expression of fungal lignocellulolytic enzymes in different heterologous hosts and the biochemical characterization of the resulting recombinant proteins.

Next-Generation Genetics in Plants: Evolutionary Tradeoffs, Immunity and Speciation

Detlef Weigel (detlef.weigel@tuebingen.mpg.de)

Max Planck Institute for Developmental Biology, Tübingen, Germany

We are addressing three core questions in evolution:

- How, and how frequently, do new genetic variants arise?
- Why do some variants increase in frequency?
- Why are some combinations of new variants incompatible?

These correspond to the fundamental evolutionary processes of mutation, selection and speciation, which we are studying using both bottom-up (i.e., forward genetic) and top-down (i.e., whole-genome) approaches.

I will begin by showcasing examples that demonstrate the power of second-generation sequencing, both in support of forward genetics (Schneeberger et al., *Nature Methods* 2009), and in determining the rate and spectrum of mutations in the plant *Arabidopsis thaliana* (Ossowski et al., *Science* 2010). Based on our experience with short-read sequencing (Ossowski et al., *Genome Research* 2008), we have been advocating a 1001 Genomes project for *A. thaliana* (Nordborg and Weigel, Nature 2009; http://1001genomes.org), and we have already sequenced 84 wild strains from this species. Interpretation of the wealth of within-species polymorphism data is greatly facilitated by outgroup information from the reference genome sequence of *Arabidopsis lyrata*, produced at JGI.

Next, I will discuss a fitness trade-off we recently discovered. The inconstancy of the environment places organisms under competing evolutionary pressures, particularly sessile organisms like plants. Allelic variants beneficial in one setting might be detrimental under different circumstances. Plants vary greatly in their ability to resist microbial or animal attack, and this is thought to reflect fitness costs in the absence of pathogens or predators. We have found that allelic diversity at a single locus, *ACCELERATED CELL DEATH 6* (*ACD6*), underpins dramatic variation in both vegetative growth and resistance to microbial infection and herbivory in *A. thaliana*. *ACD6* is also a causal factor for an autoimmune-like response that behaves as expected for Dobzhansky-Muller incompatibilities, which are often thought of underlying speciation events. Together with other findings from our group, this implicates the extreme allelic diversity of disease resistance genes (Clark et al., *Science* 2007), presumably due to pathogen pressures, as potential causes for the evolution of gene-flow barriers in plants.

Our work is supported by HFSPO, the European Commission, BMBF, DFG, and the Max Planck Society.

1. Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., and Weigel, D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. **Genome Res.** *18*, 2024-2033.

2. Schneeberger, K., Ossowski, K., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. H., Jørgensen, J.-H., Weigel, D., and Andersen, S. U. (2009) SHOREmap: mapping and mutation identification in one step by deep sequencing. **Nat. Methods** *6*, 550-551.

3. Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes. **Genome Biol.** *10*, R98.

4. Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. **Science** *327*, 92-94.

Marine Algal Evolution, Ecology and Roles in Global CO_2 Uptake

Alexandra Worden (azworden@mbari.org)

Monterey Bay Aquarium Research Institute, Moss Landing, California

The photosynthetic plankton responsible for primary production, conversion of CO_2 to organic biomass, come from an array of evolutionary histories. Marine phytoplankton are responsible for half the global primary production, which is partitioned equally between marine and terrestrial ecosystems, each accounting for approximately 50 gigatons of carbon per year. Climate change perturbations are expected to have major impacts on marine ecosystems and the biogeochemical transformations they mediate. This places tremendous urgency on gaining mechanistic understanding of biogeochemical cycles and how global CO_2 uptake by phytoplankton will transition during climate change. *Micromonas* is a tiny green alga found from the tropics to polar waters. In 2006 analysis of marker genes indicated the species designation *Micromonas pusilla* likely harbored cryptic species. However, the ecological and genomic significance of 'cryptic species' is not well understood. Complete genome analyses of two *Micromonas* clades, as well as their relative *Ostreococcus*, the smallest free living eukaryote known, reveal extensive differences, while also identifying features that appear to be ancestral – shared with land plants – but not other green algae. We have also developed a cultivation independent approach for genomic sequencing of eukaryotic algae. Here the relative distribution of both cultivated and uncultivated algae will be discussed in the context of genomic features and differentiation. From an evolutionary perspective, the genomes show cosmopolitan gene repertoires. This raises the question of whether mosaic genomes are a unique feature of a few cultured protistan taxa or a unifying feature of successful algae in modern oceans.

Flow cytometry has played a major role in oceanographic science, leading to the discovery of the most abundant photosynthetic organism on the planet, the marine cyanobacterium *Prochlorococcus*. At-sea cytometry work has also shown that tiny eukaryotes known as 'picoeukaryotes' contribute significantly to primary production. Despite their importance to global CO_2 fixation, the ecological roles and potential physiological capabilities of some phytoplankton groups, particularly picoeukaryotes, are poorly understood. The current lack of understanding is in part due to the fact that genome sequences are not available for the

majority of taxa. Some groups have yet to be cultured – excluding opportunities for laboratory experimentation or genomic analyses. Even approaches such as metagenomics, in which large volumes of seawater are filtered and DNA is sequenced directly from the environment, are of limited utility for studying eukaryotic phytoplankton. In addition to laboratory based studies on cultured species, we have been developing new approaches for genomic analysis of uncultivated unicellular eukaryotes and for quantifying their contributions to marine photosynthesis. In this talk phytoplankton diversity, and the challenges it raises from a research perspective, will be discussed, along with inroads being made on respective global contributions. In addition, genomic approaches to understanding picoeukaryote ecology and evolution will be addressed—with at-sea flow cytometry playing a central role in unraveling the mysteries of uncultivated phytoplankton.

In-silico Secretome Analysis of *Pleurotus ostreatus*

Manuel Alfaro, José L. Lavín, **Lucía Ramírez*** (lramirez@unavarra.es), José A. Oguiza, and Antonio G. Pisabarro

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

Most eukaryotic proteins are synthesized in the cytosol, and many need to be sorted to different subcellular locations. Extracellular proteins contain an N-terminal targeting sequence that is recognized by the secretory pathway, and these signal peptides (SPs) are responsible for targeting proteins to the endoplasmic reticulum for subsequent transport through the secretory pathway. In most cases, SPs are cleaved off by specific signal peptidases. Extracellular proteins in the lignocellulose degrading basidiomycete *P. ostreatus*, include the cell wall proteins (CWPs) and fully secreted enzymes. Secreted or extracellular proteins are found in the *in vitro* growth medium, and to reach the outer surface of the organisms, these proteins have to travel through the cell wall. Next to the growth medium, the cell wall is also regarded as an extra-cellular entity. *P. ostreatus* the oyster mushroom, is an active lignin degrader in the forests. Lignin is the second most abundant biopolymer on Earth and its breakdown is a necessary step for making cellulose (the most abundant carbon biopolymer) accessible to further enzymatic processes. The understanding of the whole-genome regulation of *P. ostreatus* lignocellulolytic enzymes would facilitate its use in in-situ bioremediation processes and other biotechnological processes. Recently the U.S. Department of Energy Joint Genome Institute (JGI) in cooperation with many international research groups coordinated by the Genetics and Microbiology Research Group from the public University of Navarre, has completed genome sequence of two haplotypes of the basidiomycete fungus *P. ostreatus (PC15 and PC9)*.

To perform the in-silico secretome analysis, location of the predicted presence of a N-terminal secretory pathway signal peptide was carried out using TargetP v1.1. Sequences considered as positives on the previous step, were inspected using SignalP v 3.0 to detect the presence and location of signal peptide cleavage sites in amino acid sequences and a signal peptide prediction. Transmembrane domains (TMD) were screened for each of the proteins considered positives from the previous analysis, by TMHMM v 2.0. Those proteins presenting 1 or no predicted TMD were considered positive and then extracted. A comparative genomic analysis was carried out against other phylogenetically related basidiomycete proteomes in order to assign a putative function for each of the selected proteins. Orthology relationships were determined by BLASTP and tBLASTn, based on the reciprocal best hits of each proteome against the others. Some proteins of the resulting set corresponded to conserved proteins marked as Hypothetical Protein (HP) in the databases, thus reflecting limitations in the currently available genome annotations. After the whole proteome was scanned with the previously mentioned tools, candidate proteins were extracted. A set of 463 proteins was obtained: 254 protein models with a defined function, 109 models with no defined function but appearing as conserved proteins in other fungi (HP) and 100 considered bad predicted models.

Effect of Multiple Displacement Amplification (MDA) on Sequence-Based Microbial Community Analysis

Martin Allgaier[1,2]* (MAllgaier@lbl.gov), Suzan Yilmaz,[1] and Philip Hugenholtz[1]

[1]Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, California and [2]Deconstruction Division, Joint BioEnergy Institute, Emeryville, California

DNA concentration is often a limiting factor for *omic* characterization of low-biomass environments. Multiple displacement amplification (MDA) using phi29 DNA polymerase is being used increasingly to overcome this hurdle. Despite its enormous potential, MDA has a number of recognized drawbacks, most notably amplification bias documented in single cells, which may compromise any subsequent quantitative analysis.

The impact of MDA bias on communities of microorganisms is still unknown. To address this question we performed MDA on environmental DNA samples extracted from garden compost, activated sludge and termite hindgut and assessed amplification bias using high-throughput SSU rRNA amplicon pyrosequencing.

Between 4-22% of all taxa detected were significantly skewed using MDA by either being increased or decreased in relative abundance. Observed skewing of phylotypes was not predictable based on GC content or genome size even though technical reproducibility was very high for all samples analyzed. Dominant populations were biased more frequently than low abundant members of the communities. This impact was of greater significance on less complex communities (e.g. termite gut) compared to more diverse samples (e.g. garden compost). While dominant populations were skewed around 2.5-fold in average, several taxa showed skewing of more than 20-fold. Such biases can greatly impact studies on natural microbial communities using MDA amplified DNA samples particularly if quantitative analysis (e.g. gene-centric analysis) is to be performed.

Comparison of Microbial Communities Associated with Leaf-Cutter Ant Fungus Gardens

Frank O. Aylward[1,2]* (faylward@wisc.edu), Garret Suen,[1,2] Jarrod J. Scott,[1,2,3] Sandra M. Adams,[1,2] Susannah G. Tringe,[4] Adrián A. Pinto-Tomás,[5,6] Clifton E. Foster,[1] Markus Pauly,[7] Paul J. Weimer,[8] Kerrie Barry,[4] Lynne A. Goodwin,[9] Pascal Bouffard,[10] Lewyn Li,[10] Jolene Osterberger,[11] Timothy T. Harkins,[11] Steven C. Slater,[1] Timothy J. Donohue,[1,2] and Cameron R. Currie[1,2,3]

[1]DOE Great Lakes Bioenergy Research Center and [2]Department of Bacteriology, University of Wisconsin, Madison; [3]Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama; [4]DOE Joint Genome Institute, Walnut Creek, California; [5]Departamento de Bioquímica, Facultad de Medicina and [6]Centro de Investigaciones en Estructuras Microscópicas, Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio, San Pedro de Montes de Oca, San José, Costa Rica; [7]Department of Energy (DOE) Plant Research Laboratory, Michigan State University, East Lansing; [8]Dairy Forage Research Center, U.S. Department of Agriculture–Agricultural Research Services (USDA–ARS), Madison, Wisconsin; [9]DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico; [10]454 Life Sciences, a Roche Company, Branford, Connecticut; and [11]Roche Diagnostics, Roche Applied Science, Indianapolis, Indiana

Leaf-cutter ants are dominant herbivores in the Neotropics, with individual nests capable of foraging up to 400kg dry weight plant biomass per year. These ants cultivate a mutualistic fungus on this plant material, and nutrient-rich hyphal swellings harvested from these specialized fungus gardens nourishes the ant colony. Here we present on metagenomic characterization of the bacterial community associated with leaf-cutter ant

fungus gardens and show that, as hypothesized previously, bacterial members are likely involved in important community processes such as plant biomass degradation and nitrogen fixation. Moreover, the communities associated with different leaf-cutter ant species are remarkably similar, indicating that they may be a highly conserved aspect of this system.

Structural Characterization of New Protein Families Identified from the Environment, Metagenomes, and the GEBA Project

G. Babnigg[1]* (gbabnigg@anl.gov), H. An,[1] J. Bearden,[1] L. Bigelow,[1] A. Binkowski,[1] K. Buck,[1] C. Chang,[1] G. Chhor,[1] S. Clancy,[1] M. Cuff,[1] M. Donnelly,[1] W. Eschendfeldt,[1] Y. Fan,[1] B. Feldman,[1] M. Gu,[1] C. Hatzos,[1] R. Jedrzejczak,[1] G. Joachimiak,[1] Y. Kim,[1] H. Li,[1] E. Marland,[1] B. Nocek,[1] J. Osipiuk,[1] E. Rakowski,[1] M. Schiffer,[1] A. Stein,[1] L. Stols,[1] K. Tan,[1] C. Tesar,[1] A. Weger,[1] R. Wu,[1] R. Zhang,[1] J. Thornton,[2] R. Laskowski,[2] J. Watson,[2] W. Anderson,[3] O. Kiryukhima,[3] D. Miller,[3] G. Minasov,[3] L. Shuvalova,[3] Y. Tang,[3] X. Yang,[3] C. Orengo,[4] D. Lee,[4] R. Marsden,[4] Z. Otwinowski,[5] D. Borek,[5] A. Kudlicki,[5] A.Q. Mei,[5] M. Rowicka,[5] A. Edwards,[6] E. Evdolkimova,[6] J. Guthrie,[6] A. Khachatryan,[6] M. Kudrytska,[6] A. Savchenko,[6] T. Skarina,[6] X. Xu,[6] W. Minor,[7] M. Chruszcz,[7] M. Cymborowski,[7] M. Grabowski,[7] P. Lasota,[7] P. Miles,[7] M. Zimmerman,[7] H. Zheng,[7] D. Fremont,[8] T. Brett,[8] C. Nelson,[8] C.A. Kerfeld,[9] and **A. Joachimiak**[1]

Midwest Center for Structural Genomics: [1]Argonne National Laboratory, Biosciences Division, Argonne, Illinois; [2]European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom; [3]Northwestern University, Evanston, Illinois; [4]University College London, United Kingdom; [5]University of Texas Southwestern Medical Center, Dallas; [6]University of Toronto, Ontario, Canada; [7]University of Virginia, Charlottesville; [8]Washington University, St. Louis, Missouri; and [9]DOE Joint Genome Institute, Walnut Creek, California

The Midwest Center for Structural Genomics (MCSG) as part of the Protein Structure Initiative (PSI) provides structural coverage of major protein superfamilies with granularity allowing 3D homology modeling of a large number of proteins using only computational methods. The ultimate goal of PSI is to build a foundation for 21st century structural biology where the structures of virtually all proteins will be found in the Protein Data Bank (PDB) or derived by computational methods. As part of a pilot project the MCSG used the genome sequence information generated by the Joint Genome Institute (JGI) to select protein families for structure determination. The sequenced genomes include species affecting global carbon cycling, microbial communities or single species that play a role in the degradation of lignocellulosic material, species with rich metabolic potential, as well as highly diverse microbial species sequenced as a part of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. The targets also included proteins nominated by the JGI user community. The functional characterization of some of the newly uncovered gene families, especially those without significant sequence homology to existing genes, is one of the roadblocks to modern high-throughput science. A 3D structure does not require any *a priori* knowledge about the protein, contributes new data and can jumpstart functional assignment and assay development by providing an atomic level 3D description of a given protein along with the different gene expression clones and purified proteins.

The structural genomics high-throughput pipeline developed at MCSG comprises: (1) classifying all available genomic sequences to establish a prioritized target set of proteins, (2) cloning and expressing proteins of microbial and eukaryotic origin, (3) purifying and crystallizing native and derivatized proteins for X-ray crystallography, (4) collecting data and determining structures using synchrotron sources, (5) analyzing

structures for fold and function assignment, and homology modeling of related proteins. The structural genomics pipeline takes advantage of significant advances in molecular and structural biology including synchrotron facilities, dedicated PX beamlines, advanced software and computing resources. The structural genomics technologies can be applied to a wide range of protein targets and are very well suited for proteins originating from microbial communities. The MCSG targets are selected from large protein families with no structural representative, biomedically important pathogens and higher eukaryotes, metagenomics projects and also include the scientific community nominated targets.

The MCSG started this pilot project in 2008. The JGI scientists and the User Community nominated a large set of targets from more than 50 microbial species. Nearly 70 scientists proposed more than 700 targets for structure determination. So far we have purified more than 120 targets and determined 10 structures. This poster will summarize the results of the targets processed to date with special emphasis on structures solved in the second year of this project.

http://www.mcsg.anl.gov

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

Comparative Transcriptome Analysis of Dinoflagellate Symbionts from Reef-Building Corals (*Symbiodinium* sp.)

Till Bayer,[1] Shinichi Sunagawa,[2] Mickey DeSalvo,[2] Erika Lindquist,[3] Jodi Schwarz,[4] Mary Alice Coffroth,[5] Christian R. Voolstra[1]* (christian.voolstra@kaust.edu.sa), and Mónica Medina[2]

[1]Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia; [2]School of Natural Sciences, University of California, Merced; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Vassar College, Poughkeepsie, New York; and [5]State University of New York, Buffalo

Reef building corals provide the basis of coral reef ecosystems, one of the most diverse on the planet. While corals are heterotrophs, they rely on a symbiotic relationship with dinoflagellate algae to utilitze sunlight as an energy source. Understanding the interactions between corals and their symbionts holds the key insights into this mutualistic partnership. However, thus far genetic information on the symbiont side has been limited. Funded by the Joint Genome Institute's Community Sequencing Program, transcriptome data for two strains of the highly diverged *Symbiodinium* genus are being generated via next-gen sequencing. The analysis will give insight into symbiont biology by uncovering novel genes and describing metabolic aspects through the breakdown of ESTs associated with certain pathways. Furthermore, orthologs between both strains allow for the investigation of evolutionary characteristics such as dN/dS. which can give insight into the biology of *Symbiodinium*. Specifically, rapidly evolving genes uncovered by such methods are prime candidates for proteins that may confer adaptive phenotypic traits unique to these species and their endosymbiotic lifestyle.

Mats, Microbialites and Mud: Biodiversity of Great Salt Lake Beyond Pink Water

Laura L. Beer[1]* (laurabeer@gmail.com), Chuck Pepe-Raney,[1] Natasha Zvenigorodsky,[2] Kathleen Nicoll,[3] Jonathan E. Meuser,[1] Maria Ghirardi,[4] Matthew C. Posewitz,[5] Bonnie K. Baxter,[6] and John R. Spear[1]

[1]Environmental Science and Engineering Division, Colorado School of Mines, Golden; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Department of Geography, University of Utah, Salt Lake City; [4]Biosciences Center, National Renewable Energy Laboratory, Golden, Colorado; [5]Department of Chemistry and Geochemistry, Colorado School of Mines; and [6]Department of Biology, Westminster College, Salt Lake City, Utah

This Laboratory Sequencing Project (LSP) aims to describe the microbial diversity and metabolic capabilities present in the hypersaline Great Salt Lake. Extreme salinity requires metabolic and enzymatic adaptations to tolerate osmotic stress, decreased oxygen solubility, and variations in ion availability, characteristics that may be beneficial in renewable bioenergy applications. We describe the microbial diversity of the GSL based on barcoded 454 Pyrosequencing data from 36 sites using universal primers for the V9 region of the ssu rRNA gene. Samples were collected during 2007-2009 from sites within the lake perimeter including water, benthos, microbialites and from features in the surrounding basin including springs, microbial mats, and various sediments. GPS coordinates were recorded for all sites and metadata such as temperature, salinity, dissolved oxygen, pH, ICP analyses and light intensities were obtained when possible. The GSL LSP samples were one of the first at JGI that utilized the Titanium platform adapted for amplicon sequencing and yielded over 200,000 sequence reads with the majority over 450 bases in length and 99% with quality scores greater than 27 (Q>27). The sequences were trimmed and clustered using the JGI Pyrotagger pipleline, and clustered at the OTU level using the 97% threshold. Representative OTUs were aligned using mothur, and a tree was constructed in ARB using the SILVA 3 Domain alignment. Rarefaction curves (OTUs at 3% uniqueness vs number of reads) leveled off for the hypersaline (>25% salinity) north arm and showed the least amount of richness. Consistent with previous studies, the north arm waters were dominated by members of the Halobacteria. The highest levels of species diversity were observed in mats, biofilm, sediments, and in the water from the south arm, in decreasing order. The Fast UniFrac web interface was used to examine β-diversity and the environmental factors that control microbial community composition. Initial principal coordinate analysis revealed that salinity is the main determinant of community composition. However, when the north arm water samples were removed from the analysis, the major environmental determinant of β-diversity was whether they were from water or sediments (non-water). Moreover, sediments (including mats and microbialites) showed high species (α) diversity that does not appear to be influenced by salinity. A biofilm sample yielded several alveolates, many of which are considered to be “large” members of the Ciliata, indicating that the primers also capture the microbial eukaryotes. Environmental factors influencing phylogenetic diversity and microbial evolution in this extreme environment will also be discussed.

Insertional Mutagenesis of *Brachypodium distachyon*

Jennifer Bragg[1]* (Jennifer.Bragg@ars.usda.gov), Jiajie Wu,[1,2] 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

Brachypodium distachyon (Brachypodium) is currently being developed as a model to study temperate cereals and forage grasses. It is also an ideal system for studying the basic biology underlying the traits that control the utility of grasses as energy crops. Several important genomics resources have been developed in Brachypodium including: a whole genome sequence, ESTs, SNP markers, a high-density genetic linkage map, and germplasm resources. High-efficiency transformation of Brachypodium using *Agrobacterium tumefaciens* also has been developed to facilitate functional genomic research (average efficiency 44%). The objective of our work is to generate >7,500 T-DNA insertional mutant lines, sequence the regions flanking >6,000 insertion sites, and organize this data in a searchable website to provide researchers with a means to order T-DNA lines with mutations in genes of interest. To date, we have generated >7000 T_0 lines. We have generated 3,466 flanking sequence tags (FSTs) from 4,166 of these plants. Of these FSTs, 1,601 (46.2%) contain Brachypodium genomic sequences, and the FST loci are widely distributed across all five chromosomes. Mutant phenotypes have been identified by visual screening of T_1 plants, and near-infrared spectroscopy (NIR) of powdered stem material is being used to screen the T-DNA population for lines with altered cell wall composition. A website containing protocols for working with Brachypodium, information about the T-DNA project, and instructions for ordering is now available at http://brachypodium.pw.usda.gov.

Iron Homeostasis in Yellowstone National Park Hot Spring Microbial Communities

I. Brown[1]* (igor.i.broun@nasa.gov), S.G. Tringe,[2] H. Franklin,[5] D.A. Bryant,[3] C.G. Klatt,[4] S.A. Sarkisova,[1] and M. Guevara[5]

[1]SARD/JSC, Houston, Texas; [2]Pennsylvania State University, University Park; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Montana State University, Bozeman; and [5]NASA USRP, NASA JSC, Houston, Texas

It has been postulated that life may have originated on Earth, and possibly on Mars, in association with hydrothermal activity and high concentrations of ferrous iron. However, it is not clear how an iron-rich thermal hydrosphere could be hospitable to microbes, since reduced iron appears to stimulate oxidative stress in all domains of life and particularly in oxygenic phototrophs. Therefore, the study of microbial diversity in iron-depositing hot springs (IDHS) and the mechanisms of iron homeostasis and suppression of oxidative stress may help elucidate how Precambrian organisms could withstand the extremely high concentrations of reactive oxygen species (ROS) produced by interaction between environmental Fe^{2+} and O_2.

Proteins and clusters of orthologous groups (COGs) involved in the maintenance of Fe homeostasis found in cyanobacteria (CB) inhabiting environments with high and low [Fe] were main target of this analysis.

Preliminary results of the analysis suggest that the Chocolate Pots (CP) microbial community is heavily dominated by phototrophs from the cyanobacteria (CB), *Chloroflexi*

and *Chlorobi* phyla, while the Mushroom Spring (MS) effluent channel harbors a more diverse community in which *Chloroflexi* are the dominant phototrophs. It is speculated that CB inhabiting IDHS have an increased tolerance to both high concentrations of Fe^{2+} and ROS produced in the Fenton reaction. This hypothesis was explored via a comparative analysis of the diversity of proteins and COGs involved in Fe homesotasis and suppression of oxidative stress in the CP and MS microbiomes.

In the CP microbiome, roughly 50% of the COGs involved in Fe homeostasis and the suppression of oxidative stress fell into the bacterioferritin comigratory protein, putative bacterioferritin comigratory protein, and DNA-binding ferritin-like protein functions; in contrast, the fraction of proteins with the same functions comprised only about 30% of the COGs in the MS microbiome. The majority of these COGs fell into to the O functional category – posttranslational modification, protein turnover, and chaperones – and the proteins were additionally classified as peroxiredoxins.

Within the ABC-type transport ATP-binding component function, 45% of COGs from CP fell into the V functional category – defensive mechanisms – while only 30% were classified as such in MS. No significant differences in abundance of Fe, Fe^{2+}, and Fe^{3+} transport system proteins were observed between CP and MS.

Thus, one may speculate that microbes inhabiting IDHS as Chocolate Pots possess specific proteins or metabolic cycles which help them thrive in ROS rich environment. Further details and discussion will be presented during the meeting.

Mining *Dictyoglomus turgidum* for Enzymatically Active Carbohydrases

Phillip J. Brumm,[1] Spencer Hermanson,[1] Becky Hochstein,[2] Julie Boyum,[3] Nick Hermersmann,[3] Lynne Goodwin,[4] David Sims,[4] Krishne Gowda,[3] and David Mead[3]* (dmead@lucigen.com)

[1]C5-6 Technologies and Great Lakes Bioenergy Research Center, Middleton, Wisconsin; [2]Montana State University, Bozeman; [3]Lucigen and Great Lakes Bioenergy Research Center, Middleton, Wisconsin; and [4]DOE Joint Genome Institute, Walnut Creek, California

Dictyoglomus species represent a novel group of thermophilic, anaerobic organisms with considerable biotechnological promise; these organisms are so unique that they have been given their own genus, *Dictoglomi*. *Dictyoglomus turgidum* is an anaerobic, thermophile able to degrade a wide range of biomass components including starch, cellulose, pectin and lignin. The broad range of substrate utilization is reflected in the high percentage of CAZymes present in the genome, 2.35% of its total genes, higher than the percentage present in the cellulose-degrading, thermophilc anaerobe, *Clostridium thermocellum*, with 2.2% of its total genes. This apparent strong degradation capacity, coupled with the uniqueness of the organism, suggested that *D. turgidum* would be an excellent source of new and novel biomass-degrading enzymes. Screening of a random clone library generated from *D. turgidum* resulted in the discovery of several novel biomass-degrading enzymes with low homology to known molecules. Whole genome sequencing of the organism followed by bioinformatics-directed amplification of selected genes resulted in the recovery of additional novel enzyme molecules.

The First Draft Genome Sequence of a Gram-Positive Bacterium Isolated from a Microbial Fuel Cell: *Therminicola potens* strain JR

Kathy G. Byrne-Bailey[1]* (K.G.Byrne-Bailey@berkeley.edu), Kelly C. Wrighton,[1] Ryan A. Melnyk,[1] Terry C. Hazen,[2] and John D. Coates[1]

[1]Department of Plant and Microbiology, University of California, Berkeley and [2]Earth Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley

Thermincola potens strain JR is a Gram-positive obligate anaerobe isolated from the anode of a thermophilic microbial fuel cell (MFC), where it constituted a dominant member of the current-producing bacterial community. Strain JR coupled acetate oxidation to the reduction of external electron acceptors including MFC anodes and hydrous ferric oxide (HFO). Furthermore, this bacterium forms a biofilm on an active MFC anode, but does not produce any redox active compounds to mediate reduction of the anode indicating a direct mechanism of extracellular electron transfer.

16S rRNA gene sequence analysis identified strain JR as a Firmicutes belonging to the Peptococcaceae, in the order Clostridiales and genus *Thermincola,* sharing 99% 16S sequence identity with the two previously characterized members, *Thermincola carboxdophilia* and *Thermincola ferriacetica*. This Firmicutes member is only one of a small number of *Peptococcaceae* to be genome sequenced, and represents both the first genome sequence of an MFC isolate and a *Thermincola* species. The draft genome consisted of a single circular chromosome of approximately 3036819 bp with an average G+C content of 45.9 %. A total of 2963 protein-encoding genes were predicted and 393 (6.9 %) had no similarity to public database sequences.

This organism is potentially capable of CO/CO_2 fixation using the Wood-Ljungdahl pathway (reverse acetyl-coA pathway). Energy conservation could be achieved using a carbon monoxide dehydrogenase/hydrogenase complex. No evidence for key enzymes involved in the reverse TCA cycle or the 3-hydroxypropionate pathway were identified. Key enzymes were missing (from the draft genome) for the energy and carbohydrate pathways: Calvin cycle, Embden-Meyerhof pathway, Enter-Doudoroff pathway, glycolysis and the pentose phosphate pathways.

In comparing Pfams from strain JR to its nearest neighbors with sequenced genomes, this bacterium had a larger number of proteins with double heme (CXXCH) motifs. This finding may be significant for the physiology of strain JR, as *c*-type cytochromes play an essential role in the direct reduction of extracellular electron acceptors by Gram-negative bacteria such as *Geobacter* or *Shewanella* species. Also of note, the predicted cytochrome proteins in putative gene clusters were associated with proteins containing NHL and TPR repeats, possibly indicating involvement in biofilm attachment to electrodes or direct involvement with electron transport.

Genomic analysis will aid in the elucidation of external electron transfer mechanisms by strain JR, thereby contributing to the knowledge of extracellular respiration by Gram-positive bacteria. By comparing and contrasting these mechanisms in Gram-positive and Gram-negative organisms we hope to identify both the conserved and disparate aspects of this seminal metabolic function.

Using Pathway Tools to Define Metabolic Pathways from Annotated Genomes

Ron Caspi, Ingrid Keseler, Alexander Shearer, Carol A. Fulcher, Tomer Altman, Fred Gilham, Pallavi Kaipa, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Suzanne Paley, **Miles Trupp*** (trupp@ai.sri.com), and Peter D. Karp

Bioinformatics Research Group, SRI International, Menlo Park, California

One of the main challenges currently facing the scientific community is the engineering of organisms for improved biofuel production. Bioinformatics can address this challenge in two ways. Firstly, in order to manipulate an organism to achieve a desired function there is a need to understand its metabolic network. As more sequenced genomes of bioenergy-related organisms become available, the need arises for efficient computational tools that can reconstruct metabolic networks from sequence information. Secondly, once the native metabolic network is available, there is a need for efficient tools that can aid the metabolic engineer in designing modifications to this network to achieve a desired function.

Pathway Tools[1] is a software package written by SRI International (SRI) that is capable of producing Pathway/Genome Databases (PGDBs) for organisms with a sequenced genome. The software assists users in analyzing genomic data by predicting the metabolic network of an organism, presenting pathway and genome data in user-friendly and intuitive displays, and providing tools for missing enzyme identification, omics data analyses, and comparative studies. Scientists can download the Pathway Tools software and use it to create pathway genome databases from annotated genomes.

An integral part of the Pathway Tools software is MetaCyc, a large, multiorganism database of metabolic pathways and enzymes that is manually curated by SRI. The MetaCyc database (MetaCyc.org) is a comprehensive and freely accessible resource that currently contains more than 1,400 experimentally determined metabolic pathways from more than 1,800 organisms, curated from more than 20,000 scientific publications. It also contains extensive information on metabolic enzymes, reactions, and substrates. MetaCyc is also used as a reference by the Pathway Tools software, which utilizes its experimentally-verified data for the prediction of the metabolic potential of genomes.

We have used Pathway Tools software to generate BioCyc (BioCyc.org)—a collection of more than 500 Pathway/Genome Databases, the vast majority of which are bacteria. BioCyc databases integrate the genome sequence with the predicted metabolic network, including metabolic pathways, assigned enzymes, chemical compounds, reactions, transporters, and predicted transcription units, via a user-friendly graphical interface. In addition, the Pathway Tools software contains an advanced genome browser and many tools for identifying missing enzymes, curation of additional information into the database, comparative pathway and genome analyses, and analysis of *Omics* data on global diagrams that display the full metabolic network or genome of the organism in compact designs. Pathway Tools software and database downloads are freely available to academic users.

Together, Pathway Tools and MetaCyc can be used to predict energy generating or utilizing metabolic pathways from an annotated genome and for comparative analysis of isolated strains. By searching for metabolites within predicted pathways of an annotated genome, researchers can identify necessary inputs, define excreted compounds and elucidate genes and proteins to be eliminated or added to achieve novel products or production levels.

1. Karp, P.D., et al. (2010) ``Pathway Tools version 13.0: Integrated Software for Pathway/Genome Informatics and Systems Biology," *Briefings in Bioinformatics* 11:40-79.

Microbial Metagenome Browser at UCSC

Patricia P. Chan* (pchan@soe.ucsc.edu) and Todd M. Lowe

Biomolecular Engineering, University of California, Santa Cruz

The feature-rich UCSC Genome Browser,[1] created originally to annotate the human genome, has become an established graphical tool for genome analysis for higher eukaryotes. Automated tools for new genome integration enabled us to create the UCSC Microbial Genome Browser,[2] currently with 269 bacterial and 74 archaeal genomes, available at http://microbes.ucsc.edu. As more environmental samples are sequenced, standard visualization and analysis tools are needed to study the functional characteristics and diversity within these metagenomes. We are therefore extending our microbial genome browser functionality to include metagenomic analysis. Besides the display of typical data tracks such as G/C content, protein and non-coding RNA gene predictions, and conserved protein domain matches, new features have been designed specifically for studying metagenomes. Users can identify potentially missing genes and frameshifts through BLASTX searches against reference microbial genomes. A statistical scoring model of phylogenetic gene placement enables graphical comparison of gene's phylogenetic profile within a scaffold and across the metagenome. Users can also examine a ranked list of orthologous genes by copy number within the metagenome to estimate gene or pathway frequency within the sampled community. To facilitate greater collaboration and exchange within microbial research communities, we encourage submission of external experimental and bioinformatic analyses, as well as suggestions for new analytic features.

1. Rhead, B., et al., *The UCSC Genome Browser database: update 2010.* Nucleic Acids Res, 2010. 38(Database issue): p. D613-9.

2. Schneider, K.L., et al., *The UCSC Archaeal Genome Browser.* Nucleic Acids Res, 2006. 34(Database issue): p. D407-10.

Ultra-High Resolution of Microbial Community Structures in the Human Distal Intestine

Marcus J. Claesson[1,2]* (mclaesson@bioinfo.ucc.ie), Qiong Wang,[2] Orla O'Sullivan,[3] Rachel Diniz-Greene,[1] James R. Cole,[2] R. Paul Ross,[2,4] and Paul W. O'Toole[1,2]

[1]Department of Microbiology and [2]Alimentary Pharmabiotic Centre, University College Cork, Ireland; [3]Center for Microbial Ecology and Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing; and [4]Teagasc, Moorepark Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

Variations in the composition of the human intestinal microbiota are linked to diverse health conditions. High-throughput molecular technologies have recently elucidated microbial community structures at very high resolution. Now, recent improvements in next-generation sequencing technologies allow for even deeper explorations of microbial communities than ever before. Here, we apply both longer (454 Titanium) and shorter, but more numerous, paired-end reads (Illumina) on six combinations of variable 16S rRNA

tandem regions from a single human faecal sample, in order to investigate their explorative limitations and potentials.

Based on *in silico* evaluations the V3V4 and V4V5 regions showed the highest accuracies for both technologies. However, actual sequencing of the former region revealed significant PCR bias compared to the other regions, thereby emphasising the necessity for careful primer selection. Both pyrosequencing and Illumina technologies offered higher resolution compared to their previous versions, and showed relatively consistent results with each other. Although, more than ¾ of the Illumina reads could not be classified down to genus level due to their shorter length and higher error rates over 60bp. Nonetheless, with improved quality and longer reads the massive coverage of Illumina especially will ultimately provide unprecedented insights into highly and extremely diverse environments like the human gut.

Measuring Differential Gene Expression in Caribbean Corals Exhibiting Signs of Yellow Band Disease

Collin J. Closek[1]* (cclosek@ucmerced.edu), Michael Desalvo,[1] Shinichi Sunagawa,[1] Christian R. Voolstra,[2] and Mónica Medina[1]

[1]School of Natural Sciences, University of California, Merced and [2]Marine Science and Engineering, King Abdullah University of Science and Technology (KAUST), Jeddah, Saudi Arabia

It is understood that corals have symbiotic algae and associated microbial organisms, which comprise the coral holobiont. As an integral part of the coral holobiont, prokaryotic organisms are used as indicators of coral health. Coral reefs are threatened throughout the world. These intensified threats have paralleled disease outbreaks in many endangered reefs. Yellow Band Disease (YBD) is increasingly more common and is known to lethally affect four species of boulder star (*Montastraea* spp.), as well as boulder brain coral(s) *Colpophyllia natans*. YBD epizootic events occur most prevalently in the Caribbean Sea, where *Montastraea* spp. serve as a dominant reef building species. In this study, differential transcriptomic responses of YBD corals were evaluated using newly constructed *M. faveolata* cDNA microarrays, which contain more than 11,000 features from EST data generated by JGI. Both diseased and healthy samples were competitively hybridized to microarray chips and the differential gene expression of the coral host response to YBD was measured. Using this high-throughput platform, relative gene expression between healthy and diseased colonies highlighted genes that are potential indicators of YBD onset.

The Labyrinthulomycetes: Little Known Marine Fungus-Like Protists Contribute to Remineralization of Organic Material and Trophic Upgrading of Poor-Quality Detritus

Jackie L. Collier* (jcollier@notes.cc.sunysb.edu), Kirk Apt, Daiske Honda, Celeste Leander, David Porter, Seshagiri Raghukumar, and Clement Tsui

School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York

The labyrinthulomycetes are ubiquitous, diverse, and abundant marine protists. They are thought to live mainly as saprobes, obtaining their nutrition from non-living particulate organic matter (POM) of algal, higher plant, or animal origin. Thus, while labyrinthulomycetes are not fungi in a taxonomic sense, they function as fungi in an ecological sense, likely playing important roles in the decomposition of marine POM. They are capable of utilizing a wide range of biopolymers, including relatively refractory substrates such as lignocellulose and sporopollenin, and produce a wide variety of hydrolytic enzymes including cellulase, xylanase, proteases, esterases, lipases, and phosphatases. Because of their high content of essential long-chain polyunsaturated fatty acids (LCPUFAs), and ability to synthesize LCPUFAs *de novo*, labyrinthulomycetes may also play a role in the nutrition of marine metazoans by improving the food quality of detritus ('trophic upgrading'). Their production of LCPUFAs has also made them organisms of increasing biotechnological interest; in fact, LCPUFAs produced by labyrinthulomycetes are already marketed as human food supplements. Labyrinthulomycetes have additional biotechnological potential in production of carotenoids, sterols, and other metabolites, and in the conversion of low-quality biomass into high-value products. The diversity of labyrinthulomycetes in marine ecosystems is much greater than the strains currently available in culture, suggesting that much remains to be learned about their ecology and physiology. As heterotrophic members of the heterokont lineage, which includes a variety of photoautotrophs, they also offer a window into the evolutionary history of plastids in major eukaryotic groups.

Considering their importance in marine carbon cycling, their biotechnological potential, and their unique evolutionary position, the labyrinthulomycetes are a vastly under-studied group of organisms. The genomes of four labyrinthulomycetes, as well as 50,000 to 100,000 ESTs for each genome to support gene identification and annotation, will be sequenced. The species chosen are *Aurantiochytrium* (formerly *Schizochytrium*) *limacinum* ATCC MYA-1381, *Schizochytrium aggregatum* ATCC 28209, *Aplanochytrium kerguelense*, and *Labyrinthula terrestris* ATCC MYA-3074. Assuming that the available pulsed-field gel electrophoresis genome size estimates are representative of most labyrinthulomycetes, but are likely to be substantial underestimates of true genome size, each genome may be approximately 20 Mb. This project will bring labyrinthulomycetes quickly into reach of modern molecular genetic methods by producing the annotated genome sequences of four labyrinthulomycete species chosen to represent important aspects of the known physiological and taxonomic diversity of the group. Comparative analysis of the genome sequence data will provide insights into their ecological roles in the cycling of carbon and other elements, and open new opportunities for their biotechnological use by revealing details of both known and novel metabolic pathways.

Directed Evolution of Ionizing Radiation Resistance in *Escherichia coli*

Michael M. Cox[1] and **John R. Battista[2]*** (jbattis@lsu.edu)

[1]Department of Biochemistry, University of Wisconsin, Madison and [2]Department of Biological Sciences, Louisiana State University and A&M College, Baton Rouge

There are several dozen known bacterial species that display an extraordinary resistance to the effects of ionizing radiation (IR). The best studied of these is *Deinococcus radiodurans.* Although the literature is replete with well-constructed hypotheses, the molecular basis of the extreme IR resistance of *Deinococcus* is not well understood. The evolution of IR resistance has been linked to the evolution of desiccation tolerance. By evolving IR resistance in a relatively sensitive bacterial species, we have tried to focus on key mechanisms of IR resistance as illustrated by the genomic changes in the evolved strains. Four populations of *Escherichia coli* K12 were independently derived from MG1655, each specifically adapted to survive exposure to high doses of IR. Complete genomic sequencing was carried out on twenty-five purified strains derived from these populations. When the sequences were compared between single colony isolates recovered from each independently adapted population, the results suggested that each population evolved a separate mechanism of radioresistance; we found little evidence of overlap among functions potentially affected by mutation. After this initial analysis, we made the assumption that it would be more beneficial to compare more closely related strains – the closer two related IR-resistant strains are to each other, the more likely they are to share the same mechanisms of protection from IR. We sequenced and compared the genome of isolates obtained from a single evolved population. In addition, we determined when an allele first appears in the evolving population and establish if the frequency of that allele increases in the populations isolated after each successive application of IR using a method designed to detect single nucleotide polymorphisms in the presence of a 1000 fold excess of wild type DNA. In this way we obtain a measure of the potential significance of any allele to IR resistance. Results to date support a role for modifications in RecA and the replication restart primasome in the increased IR resistance of these strains. These results are exciting in that they greatly reduce the complexity of the problem we face in identifying critical components of ionizing radiation resistance by cleanly "separating the wheat from the chaff" among the collection of mutants identified by sequencing. The technique outlined also provides us with the opportunity to follow an evolutionary process in remarkable (possibly unprecedented) detail.

Project Management and Targeted Finishing in a High-Throughput Finishing Pipeline.

K. Davenport* (kwdavenport@lanl.gov), L. Meincke, L. Goodwin, O. Chertkov, C. Han, and C. Detter

Los Alamos National Laboratory, Los Alamos, New Mexico

New developments in high-throughput finishing include automated start-up, automated closeout, targeted finishing, and project management. With the increasing number of projects coming into our finishing pipeline we are actively developing new ways to streamline the finishing process. We're working to decrease the amount of time a project spends in manual finishing by improving and automating the first phase of genome finishing as well as implementing a formal project management system. These improvements will allow for a faster genome finishing process while reducing costs and

improving overall tracking and management of projects at JGI-LANL. Additionally, at a collaborator's request, we can quickly finish targeted regions of interest within a genome while still maintaining the capability to finish the entire genome. We have developed these new targeted finishing strategies in an effort to aid specific and important research interests of the Bioenergy Centers. Several projects have undergone this new approach which has produced the targeted data of interest in a much shorter turnaround time to aid researchers for various bioenergy studies.

Development of High Throughput Process for Constructing 454 Titanium Libraries

Shweta Deshpande[1]* (SDeshpande@lbl.gov), Eric Tang,[1] Chris Hack,[1] Susan Lucas,[2] Jan-Fang Cheng,[1] and the JGI Production Sequencing Group

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

We have developed a process with the Biomek FX robot to construct 454 titanium libraries in order to meet the increasing library demands. All modifications in the library construction steps were made to enable the adaptation of the entire processes to work with the 96-well plate format. The key modifications include the shearing of DNA with Covaris E210 and the enzymatic reaction cleaning and fragment size selection with SPRI beads and magnetic plate holders. The construction of 96 Titanium libraries takes about 8 hours from sheared DNA to ssDNA recovery. Although this process still require manual transfer of plates from robot to other work stations such as thermocyclers, these robotic processes represent about 12- to 24-folds increase of library capacity comparing to the manual processes. To enable the sequencing of many libraries in parallel, we have also developed sets of molecular barcodes. The requirements for the 454 library barcodes include 10 bases, 40-60% GC, no consecutive same base, and no less than 3 bases difference between barcodes. We have used 96 of the resulted 270 barcodes to construct libraries and pool to test the ability of accurately assigning reads to the right samples. When allowing 1 base error occurred in the 10 base barcodes, we could assign 99.6% of the total reads and 100% of them were uniquely assigned. We have begun to assess the ability to assign reads after pooling different number of libraries. We will discuss the progress and the challenges of this scale-up process.

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

Acclimation of *Symbiodinium sp.* (KB8) to High Growth Temperature

Erika Díaz-Almeyda* (ediaz-almeyda@ucmerced.edu) and Monica Medina (mmedina@ucmerced.edu)

School of Natural Sciences, University of California, Merced

Coral reefs are most biodiverse marine ecosystems on Earth and provide important services to humans such as tourism, fishing, coastal protection and pharmaceutical products, among others. The health and productivity of these ecosystems hinges on a functional symbiosis between the cnidarian hosts and photosynthetic dinoflagellate (*Symbiodinium*) endosymbionts. Coral bleaching is the result of the disruption of this symbiosis. Bleaching is often caused by thermal stress and the symbiont's photosynthetic apparatus is extremely sensitive to thermal fluctuations. In this study, we cultured *Symbiodinium sp.* (KB8) under control (i.e. ambient reef) and high temperature conditions. We found differences in several physiological parameters as growth rate, chlorophyll a and c content, membrane fluidity and electron transport rate of the PSII in the treatment comparisons. Additionally, differential gene expression was assessed by the use of a KB8 gene chip (853 ESTs) generated from ESTs sequenced by JGI. This study will be helpful in enhancing our understanding of the molecular and cellular mechanisms involved in *Symbiodinium* temperature acclimation.

High Quality Annotation of the Filamentous Fungus *Ashbya gossypii*

Fred S. Dietrich[1]* (dietrich@duke.edu), Sidney Kuo,[1] Sylvia Voegeli,[2] and Peter Philippsen[2]

[1]Institute of Genome Sciences and Policy and Molecular Genetics and Microbiology Department, Duke University, Durham, North Carolina and [2]Biozentrum, University of Basel, Basel, Switzerland

Ashbya gossypii is a filamentous hemiascomycete used commercially for the production of riboflavin. The genome sequence of this fungus was carried out in the laboratory Dr. Peter Philippsen at the Biozentrum, University of Basel and fully annotated, deposited in Genbank, and published in 2004. The initial sequence was complete other than three small gaps and four telomere sequences. This sequencing revealed that this fungus has a remarkably small genome of less than 10Mb, a gene set of only 4718 genes, and no transposable elements.

Recently we have undertaken resequencing this genome to deep coverage, sequencing the genome of a second strain *A. gossypii*, and sequencing the genome of a second *Ashbya* species, *A. aceri*, and have undertaken a thorough reannotation of the *A. gossypii* genome. By carrying out a careful in-depth annotation we have identified genes not previously identified, identified an unusual alternative splicing pattern not known in *Saccharomyces cerevisiae*, identified eight genes translated across frameshifts, identified the MATalpha genes not previously known in *A. gossypii*, identified the mechanism by which the MATalpha locus was lost in the laboratory strain, validated that there are overlapping open reading frames in this organism, and uncovered clues as to the origin of riboflavin overproduction in this strain.

In this poster we will discuss the additional benefits in biological insights gained in going from a good annotation based on a nearly complete but 99.8% accurate genome sequence

to a very careful annotation based on multiple highly accurate complete genome sequences and automated annotation followed by careful hand curation.

Metagenomic and Transcriptomic Techniques to Resolve the Process of Wood Digestion in the Asian Longhorned Beetle (*Anoplophora glabripennis)*

Scott Geib,[1] Erin Scully[2]* (eds14@psu.edu), John Carlson,[3] Ming Tien,[1] and Kelli Hoover[4]

[1]Department of Biochemistry and Molecular Biology, [2]Department of Genetics, [3]Department of Forestry, Huck Institute of the Life Sciences, and [4]Department of Entomology, Pennsylvania State University, University Park

Lignocellulose degradation in the guts of wood-feeding insects may serve as a novel source of efficient enzymes that could be exploited to degrade lignin on an industrial scale. The **A**sian **l**onghorned **b**eetle (ALB; *Anoplophora glabripennis*) feeds and grows in a harsh environment devoid of many nutritional resources by utilizing intractable components, including lignin and cellulose, for energy and protein. Our lab recently demonstrated that the lignin macromolecule is degraded during passage through the ALB gut. Through 16s rDNA and ITS amplicon sequencing, we determined that the ALB gut harbors a rich diversity of microbiota, including several genera known to play significant roles in wood decay and other bacteria that could be involved in novel digestive processes (Geib et al, 2009). Thus, it seems likely that this consortium of gut symbionts may contribute to lignocellulose digestion.

We have devoted much of our research to investigating lignocellulose degradation in ALB and the involvement of gut microbiota in these processes. Furthermore, DOE has recognized the potential of ALB to provide significant contributions to biofuels production and through our collaboration with the Joint Genome Institute, we have obtained a preliminary assembly and putative annotations for the ALB gut microbial community metagenome. MEGAN analysis of assembled contigs indicates thepotential presence of genes from many phylogenetically diverse bacterial and fungal clades including, alpha-, beta-, and gamma-proteobacteria, and ascomycota. Based on the diversity of our metagenomic sample, we have the potential to discover many novel genes that could be exploited for industrial lignocellulose digestion.

In concert, we recently sequenced the ALB gut transcriptome to determine if any novel lignin-degrading enzymes or cellulases could be identified. Based on preliminary assembly and annotation, we identified a number of contigs that share a strong degree of sequence homology with insect-derived endoglucanases and beta-glucosidases, which are involved in biochemical pathways responsible for cellulose digestion. We also identified a number of contigs that encode hydrolytic and proteolytic domains, which are likely involved in digestion; however, we have yet to discover any transcripts that could be responsible for lignin degradation. This likely reflects a lack of depth in our transcriptome, which is mirrored by the paucity of bacterial-derived contigs detected in our assembly. Therefore, we are currently working to prepare a more in-depth gut bacterial metatranscriptome in an attempt to identify bacterial genes responsible for the production of lignin degrading enzymes.

Characterizing the Microbial Community in Fiber and Epithelial Fractions of the Crop of Chick and Adult Hoatzins using 16S Pyrotags

Filipa Godoy-Vitorino[1]* (filipagodoyvitorino@gmail.com), Anna Engelbrektson,[1] Maria A. Garcia-Amado,[2] Fabian Michelangeli,[2] Philip Hugenholtz,[1] and Maria Gloria Dominguez-Bello[3]

[1]Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, California; [2]Instituto Venezolano de Investigaciones Científicas (IVIC), Caracas, Venezuela; and [3]Department of Biology, University of Puerto Rico, Rio Piedras Campus, San Juan, Puerto Rico

The hoatzin *(Opisthocomus hoazin)* is a South American strict folivorous bird, and is unique among known avian species by virtue of the fermentative function of its enlarged crop, analogous to the rumen of Artyodactyla. The crop harbors an impressive array of microorganisms with potentially novel cellulolytic enzymes. The aim of this study was to profile the microbial community in the fiber and epithelial fractions of a collection of chick and adult crops to facilitate selection of samples for metagenomic analyses, based on community stability (reproducibility) and presence of potential fiber degrading species.

Genomic DNA was extracted from crop fiber and epithelial material of 7 chicks, 1 juvenile and 7 adult birds as well as from the liquid fraction of 1 adult, with ~3 biological replicates per sample. The V8 region of the 16S rRNA gene was amplified and pyrosequenced. Reads were classified with Pyrotagger using a 10% quality filter and 97% sequence identity for clustering.

Our results demonstrate that on average the adult crop microbiota is richer (3,579$\pm$ 1000 OTUs) than that of the chicks (1,365$\pm$500 OTUs) with no differences between the epithelium or fiber fractions, and with the liquid fraction being the richest (5,290 $\pm$ 1400 OTUs).

Among non-bacterial microorganisms, Methanomicrobia dominated the chicks' crop while the juvenile and the adults were dominated by Alveolata (which includes ciliates). The epithelial fraction had a higher abundance of Methanomicrobia, followed by Alveolata and Methanobacteria. Equal proportions of Methanobacteria and Alveolata were characteristic of the fiber fraction. Fungi were present with similar abundances in both fractions and the liquid fraction was dominated by Alveolata.

Overall, the bacterial biota was dominated by Firmicutes, Bacterioidetes and Proteobacteria with lesser representation of 35 additional phyla. Chicks had more Proteobacteria and Planctomyces than the adults, but the overall phylogenetic comparisons of the microbial communitites between chicks and adults show no significant differences. The chick samples were highly non-reproducible compared to the adult samples.

Furthermore, similarity analyses (SIMPER) using the PRIMER package—aimed at identifying the OTUs that primarily discriminate between fiber and epithelial fractions—indicate that bacterial lineages that contribute to these differences include Prevotella and other Bacteroidales, which were more abundant in fiber.

These results confirm the presence of some of the rarer phyla that had been previously detected by microarray (PhyloChip) analysis, including Spirochaetes, Fibrobacteres and Chloroflexi and suggest that the fiber fraction of the adult samples will be the most interesting for metagenomic studies.

Biomineralization-Related Proteins in Metazoan Lineages

Bishoy Hanna,[1]* Christian Voolstra,[2] and Monica Medina[1]

[1]School of Natural Sciences, University of California, Merced and [2]Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Biomineralization, a process that is found across the Tree of Life, is the formation of minerals by living organisms. Animal calcification in particular is the controlled deposition of calcium carbonate salts to produce supportive structures in both vertebrate (i.e. bones, teeth) and invertebrate species (i.e., molluscan shells, coral skeletons). The ability to calcify was an evolutionary innovation that is thought to be greatly responsible for what has been coined the "Cambrian explosion", a major adaptive radiation event that took place at the late Neoproterozoic-Cambrian boundary (~545 mya). Multiple extant animal (metazoan) phyla have the ability to calcify, however, little is known about the genetics of the ancestral metazoan biomineralization toolkit. Therefore, at present it is not possible to establish with clarity whether there was an ancestral metazoan biomineralization core toolkit shared by all the extant calcifying animal lineages. Genomic and post-genomic approaches are opening a new window of opportunity to engage in questions regarding homology and evolution of this important biological innovation, as they provide both a means to analyze the overall composition of the animal biomineralization secretome, and allow for the incorporation of non-model organisms that will assist in filling the phylogenetic gap. In order to test different hypotheses related to the evolution of the biomineralization toolkit we developed a bioinformatics pipeline that we used to identify shared protein families involved in biomineralization related processes across cnidarians, mollusks, echinoderms, and vertebrates. By constructing a comprehensive list of biomineralization proteins from the scientific literature and the Gene Ontology project database AMIGO, we identified all Pfam domains associated with the candidate protein list. Using a combination of the HMMER package and BLAST in JGI sequenced whole genomes and EST datasets, we were able to look at the distribution of shared families between all these different groups. A web-based database was constructed to cater for all the data generated through the pipeline and is available at http://sequoia.ucmerced.edu/biomin/.

Development of a High-Throughput Transcriptomics (RNAseq) Analysis Pipeline

Loren Hauser[1]* (hauserlj@ornl.gov), Daniel Quest,[1] Andrey Gorin,[1] Steve Brown,[1] Sara Blumer-Schuette,[2] Mike Adams,[3] Bob Kelly,[2] and Bob Cottingham[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]North Carolina State University, Raleigh; and [3]University of Georgia, Athens

High-throughput Transcriptomics (RNAseq) has the potential to dramatically improve understanding of biological systems. RNAseq has substantially better dynamic range (10^6) and sensitivity (1 mRNA molecule per 10^3 bacterial cells) than traditional gene expression arrays (10^3). It also provides direct sequence data for determining operon structure and finding new regulatory RNAs. In order to establish the potential of High-throughput Transcriptomics (RNAseq) as an enhancement to the JGI Microbial Genome Program we have initiated the following demonstration project. We will compare the Transcriptome Profiles (TPs) of selected members of the *Caldicellulosiruptor* genus, a group of high growth-temperature biomass degraders, which have different phenotypes when grown on alternative carbon sources, and grow relatively fast on plant biomass. The TPs will

provide unique and detailed phenotypes of each strain to assess their use in consolidated bioprocessing of biomass for biofuel production. We have initiated construction of an analysis pipeline to generate TPs. These profiles, plus additional data such 5' end sequencing via RACE, cluster analysis from standard gene arrays, and Transcription Factor Binding Site (TFBS) predictions for each of the 8 species will provide the basis for building Genetic Regulatory Networks (GRNs), which will eventually enable more efficient experimental design and genome engineering. In addition, since >95% of the genes are detected when multiple growth states are analyzed and genes are typically 90% of most bacterial genomes, RNAseq data will also be useful in genome assembly and finishing.

Fungal Degradation of Cellulosic Biomass: A Transcriptomics Approach to Enzyme Discovery

Corinne D. Hausmann[1,2]* (chausmann@berkeley.edu), William Beeson,[1,2,3] and Jamie Cate[1,2,3]

[1]Energy Biosciences Institute, [2]Departments of Molecular and Cell Biology and [3]Chemistry, University of California, Berkeley

Cellulose, the most abundant organic compound on earth, is a major structural component of the primary plant cell wall. The dwindling supply of a finite amount of fossil fuels dictates the need to develop alternative energy sources to alleviate the need for petroleum. As a renewable source of energy, the degradation of plant biomass into biofuels provides an excellent alternative to crude oil. As such, much attention has been focused on improving the process of degrading biomass, with special emphasis on improving and streamlining industrial applications in an economically feasible manner. *Sporotrichum thermophile*, a filamentous fungi, secretes robust, cellulolytic enzymes in high yields that efficiently break down cellulose under industrially relevant temperatures and may prove to be an excellent source of low-cost enzymes necessary for the conversion of biomass into usable energy. Through a combination of transcriptomics and proteomics analyses, we identified suites of enzymes induced on various cellulosic substrates, providing mechanistic insight into the strategies used by *S. thermophile* to degrade lignocellulosic biomass. Further, we hope to manipulate these cellulosic enzymes, which may be useful in industrial settings for the production of biofuels.

Benchmarking Ribosomal RNA Removal Methods for Microbial Metatranscriptomics

Shaomei He[1]* (SHe@lbl.gov), Omri Wurtzel,[2] Kanwar Singh,[1] Jeff Froula,[1] Suzan Yilmaz,[1] Zhong Wang,[1] Feng Chen,[1] Erika Lindquist,[1] Rotem Sorek,[2] and Philip Hugenholtz[1]

[1]DOE Joint Genome Institute, Walnut Creek, California and [2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

The predominance of ribosomal RNA in the transcriptome is a major technical challenge in metatranscriptomics. Ribosomal RNA removal has been applied in a number of studies to enrich mRNA, yet there has not been a systematic study on their effectiveness in mRNA enrichment and bias in mRNA transcript abundance introduced by removal procedures. In this benchmarking study, we investigated the effectiveness and fidelity of several commonly used rRNA removal methods, based on subtractive hybridization (Hyb) and/or

exonuclease (Exo) digestion. These methods were evaluated on two synthetic metatranscriptomes constructed from a set of prokaryotes with sequenced genomes, using Illumina GA2 sequencing and technical replicates. We found that rRNA removal is community dependent for both Hyb and Exo treatments and that two rounds of Hyb did not significantly increase rRNA removal over one round. Organisms resistant to either Hyb or Exo were susceptible to rRNA removal when both methods were applied together. However, relative transcript abundance fidelity was significantly compromised by Hyb and Exo used in combination. Transcript fidelity was only preserved using Hyb alone. Technical replicates were highly reproducible with the exception of one set of inter-run replicates. We noted that lower run quality caused a systematic bias against high GC templates using Illumina sequencing, potentially skewing quantitative inter-run comparisons of metatranscriptomic data. We recommend applying a single round Hyb to total RNAs for metatranscriptomic analyses and to only perform intra-run comparisons for accurate quantitative comparisons of Illumina data.

Characterization of Microbial Strains Important in Biofuels Production and Biomass Conversion

C.L. Hemme[1,2]* (hemmecl@ou.edu), L. Lin,[1,2] W. Liu,[1,2] M.W. Fields,[3] Q. He,[4] Y. Deng,[1,2] Q. Tu,[1,2] H. Mouttaki,[1,2] Z. He,[1,2] K. Barry,[5] E.H. Saunders,[6] H. Sun,[6] M. Land,[7] L. Hauser,[7] A. Lapidus,[5] C.S. Han,[6] J. Wiegel,[8] R. Tanner,[2] Lee Lynd,[9] P. Lawson,[2] A. Arkin[11] C. Schadt,[12] B.S. Stevenson,[2] M. McInerney,[2] Y. Yang,[1,2] H. Dong,[13] R. Huhnke,[14] J.R. Mielenz,[12] S.-Y. Ding,[15] M. Himmel,[15] S. Taghavi,[16] D. van der Lelie,[16] E. Rubin,[5] and J. Zhou[1,2]

[1]Institute for Environmental Genomics and [2]Department of Microbiology, University of Oklahoma, Norman; [3]Montana State University, Bozeman; [4]Department of Civil and Environmental Engineering, University of Tennessee, Knoxville; [5]DOE Joint Genome Institute, Walnut Creek, California; [6]DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico; [7]Genome Analysis and Systems Modeling Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [8]Department of Microbiology, University of Georgia, Athens; [9]Dartmouth College, Hanover, New Hampshire; [11]University of California, Berkeley; [12]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [13]University of Miami, Oxford, Ohio; [14]Oklahoma State University, Stillwater; [15]National Renewable Energy Laboratory, Golden, Colorado; and [16]Brookhaven National Laboratory, Upton, New York

Genomic sequencing of 20+ clostridia strains related to biofuels production and biomass conversion were sequenced, including multiple strains from Cluster III thermophilic and mesophilic cellulolytic *Clostridium* species and multiple strains of saccharolytic *Thermoanaerobacter* species. This dataset represents a significant improvement in the genomic knowledge base of bacteria important to biofuels production. The genomes of three *Thermoanaerobacter* strains from this group, *T. pseudethanolicus* 39E, *Thermoanaerobacter* sp. X514 and *T. italicus* Ab9, have been finished and comparative genomics analysis has been conducted. Comparative analysis of the four complete genomes provides the most detailed view to date of the dynamics of *Thermoanaerobacter* genome evolution. Strains derived from hot springs environments show a conserved genome structure whereas the single subsurface strain X514 shows a highly dynamic genome that suggests possible roles of environment on shaping genomic architecture. Further analyses of the metabolic profiles of the strains resulted in testable hypotheses regarding the relative carbon uptake and usage profiles of the strains. These hypotheses were experimentally tested using metabolics and molecular biology techniques to identify physiological traits relevant to ethanol fermentation and biomass degradation in co-culture with select *Clostridium thermocellum* strains. In particular, it was observed that the three

strains employ distinct lineage-specific xylose metabolism and transport systems and that X514 shows a significantly increase absolute rate of carbon flux from xylose compared to 39E. Furthermore, it was show that the ability of X514 to synthesize vitamin B_{12} *de novo* alleviates the need to supplement ethanol-producing cultures with that vitamin, whereas ethanol yields from 39E cultures are very sensitive to B_{12} concentrations due to the inability of 39E to synthesize that vitamin. The transcriptional profiles of *Thermoanaerobacter* sp. X514 grown on different carbon substrates was also determined. Experimental studies have shown that X514 is able to metabolize hexose (glucose, fructose, galactose) and pentose (xylose and ribose) monosaccharides as well as some complex carbohydrates (cellobiose, starch and sucrose). When X514 is grown on a given substrate (ie xylose), the corresponding metabolism genes are highly expressed as expected. X514 employs both the Embden-Meyerhof-Parnas (EMP) and pentose phosphate (PPP) pathways for sugar metabolism and encodes carbohydrate active enzymes specific to fructose, xylose and cellobiose. In contrast to glucose metabolism, growth on xylose, fructose or cellobiose results in a shift in the carbon flux towards ribose, suggestions increased production of substrates for nucleotide and amino acid biosynthesis. Experimental evidence shows higher rates of energy metabolism when X514 is grown on fructose and higher yields of acetate, ethanol and lactate. Furthermore, V-type ATPase genes and a large number of genes involved in inorganic ion transport and metabolism (i.e. Na-translocating decarboxylase, Na+/H+ antiporters, etc.) are significantly upregulated, suggesting increased energy metabolism and ATP production during grown on fructose. All conserved *Thermoanaerobacter* alcohol dehydrogenase genes are expressed at similar levels on all substrates, but adh genes specific to X514 showed differential expression under different growth conditions. Finally, the accumulated genomic knowledge base of ~140 clostridia genomes was used to conduct a phylogenomics analysis of the clostridia to identify genomic relationships and properties of the clade. Analysis provides the first genomic rationale for classification of *Lachnospiraceae*, confirms previous classifications of *Clostridium sensu stricto* and strengthens a previous genomic association between Class III cellulolytic *Clostridium* and Class V saccharolytic *Thermoanaerbacteraceae* species.

Breaking the Lignocellulolytic Barrier in Biofuel Production: Discovery of Novel Cellulases from Rumen Microbes by Ultra-Deep Sequencing

M. Hess* (mhess@lbl.gov), A. Sczyrba, Z. Whang, T.W. Kim, D.S. Clark, R. Mackie, and E.M. Rubin

DOE Joint Genome Institute, Walnut Creek, California

It has become apparent that fossil fuels can only be replaced with biofuels if inexpensive and highly lignocellulolytic enzymes become available.

In the study presented here, we employed sequenced based metagenomics to identify more than 200 full-length cellulases from rumen microbes that colonized switchgrass fiber during an incubation period of 3 days. The identified full-length cellulases had an average sequence identity of 50%, suggesting that many of their characteristics are different from the characteristics of currently known cellulases. We expressed 96 of the identified candidates and verified cellulase activity for 17% of them using biochemical activity tests. Activity against solid substrates was detected as well.

The strategy developed during this project can be easily adapted for other target enzymes and will facilitate the large-scale identification of novel biocatalysts.

DNA Subway Places Students on Fast Track to Gene Annotation and Genome Analysis

Uwe Hilgert* (hilgert@cshl.edu), Cornel Ghiban, Eun-Sook Jeong, and David Micklos

iPlant Collaborative/Dolan DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

As genomic data increasingly drive biological research, biology education needs to integrate data and bioinformatics tools that originally were designed for expert users. ***DNA Subway*** (http://www.dnasubway.org) places plant genomes into the hands of students and faculty, following the example of websites such as *BioServers* and *GeneBoy* to present complex scientific tools in an intuitive and interesting interface while maintaining their scientific validity.

Annotating and comparing genome sequences can bring to life elements of gene structure and function that previously could only be approached as abstractions. Assembling gene models and comparing genes engages students in their own learning and *DNA Subway* provides the collaborative workspace to do so; offering facilities to create projects, upload or access DNA contigs, and analyze genomic data in streamlined bioinformatics workflows.

Using the metaphor of subway lines to present bioinformatics workflows lowers potential entrance barriers for users new to analyzing genomic data, easing confusion and apprehension. "Riding" any of three different lines of the *DNA Subway*, users can analyze up to 10 megabases of DNA – predicting and annotating genes (Red Line), prospecting genomes for related genes (Yellow Line), and analyzing next-generation transcriptome data (Blue Line). Through algorithms installed on *iPlant Collaborative* servers, users 1) predict protein-coding genes and tRNAs; 2) BLAST- and BLAT-search customized UniGene, UniProt, and genomic databases; 3) construct and edit gene models in a graphical annotation editor (*Apollo*); 4) view assembled data in a stand-alone browser; and 5) export results to Phytozome and other community browsers to view them in the context of completed and annotated genomes. Projects can be saved to a personal profile and shared with other users.

Metagenomic Sequencing of High-Temperature Microbial Communities in Yellowstone National Park: Phylogenetic and Functional Analysis

W. Inskeep[1]* (binskeep@montana.edu), **Z. Jay,**[1] and **YNP Metagenome Project Participants:** K. Barry,[2] S. Boomer,[3] E. Boyd,[1] I. Brown,[4] D. Bryant,[5] B. Fouke,[6] N. Hamamura,[7] M. Herrgard,[8] C. Klatt,[1] M. Kozubal,[1] A. Lapidus,[2] S. Lowry,[2] T. McDermott,[1] D. Mead,[9] S. Miller,[10] D. Rusch,[11] N. Parenteau,[12] A.-L. Reysenbach,[13] F. Roberto,[14] A. Schwartz,[8] J. Spear,[15] C. Takacs-Vesbach,[16] S. Tringe,[2] W. Ward,[1] and M. Young[1]

[1]Montana State University, Bozeman; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Western Oregon University, Monmouth; [4]Johnson Space Center, NASA, Houston, Texas; [5]Pennsylvania State University, University Park; [6]University of Illinois, Urbana; [7]Ehime University, Matsuyama, Japan; [8]Synthetic Genomics Inc., La Jolla, California; [9]Lucigen Corporation, Middleton, Wisconsin; [10]University of Montana, Missoula; [11]J. Craig Venter Institute, Rockville, Maryland; [12]NASA Ames Research Center, Mountain View, California; [13]Portland State University, Portland, Oregon; [14]Idaho National Laboratory, Idaho Falls; [15]Colorado School of Mines, Golden; and [16]University of New Mexico, Albuquerque

The *Yellowstone Metagenome Project* is a collaborative effort that was catalyzed in part through networking activities of the NSF Research Coordination Network focused on *Geothermal Biology and Geochemistry in Yellowstone National Park (YNP)*. The primary goal of this DOE-JGI Community Sequencing Project (2007-2008) was to acquire and analyze a comprehensive metagenomic dataset of the diverse thermophilic prokaryotic communities inhabiting geochemically distinct geothermal sites within the Yellowstone geothermal complex. Twenty geothermal sites were selected to achieve a representative range of geochemical and physical conditions common in YNP (www.rcn.montana.edu), and to capitalize on an extensive foundation of prior (and on-going) research and characterization of these low-diversity microbial systems.

Twenty geothermal microbial mats and or sediments were sampled in 2007 and 2008 and used to extract DNA of sufficient quantity and quality for construction of small insert libraries, which were subjected to Sanger sequencing at DOE-JGI (4 sites subjected to pyrosequencing). Aqueous and solid phase samples were also collected for analysis of predominant geochemical constituents and mineralogy associated with each microbial community. Analysis of random shotgun sequence data from chemotrophic environments (65-88 C) reveal dominant archaeal populations within the Crenarchaeota (e.g, orders Sulfolobales, Thermoproteales, Desulfurococcales), and sub-dominant populations within the Euryarchaeota as well as Candidate Phylum Thaumarchaeota. The functional attributes of gene sequence attributable to these phyla suggest a strong correlation between organism distribution and environmental parameters such as temperature, pH, dissolved sulfide and dissolved oxygen. The predominant bacteria within the chemotrophic habitats included in this study are members of the order Aquificales, and significant diversity within this group exists both within and across different geothermal systems. Other less-dominant bacterial species noted in high-temperature chemotrophic systems include members of the Proteobacteria, Firmicutes, Deinococcus-Thermus, Thermotogae, and other uncharacterized groups. The diverse phototrophic systems represented in the study include both oxygenic and anoxygenic mats ranging in temperature from 48-66 C, where members of the Chloroflexi and Cyanobacteria are common community members across highly diverse environments.

The YNP Working Group met preceding the 2009 JGI User's Meeting to discuss data analysis approaches, data interpretation, IMG utilities, and to outline four manuscripts intended to synthesize phylogenetic and functional interpretations of indigenous genomic content across different thermophilic habitats in YNP. The goal of these manuscripts, which are now in preparation, is to summarize and interpret the metagenome data as a function of environmental parameters. The metagenome sequence data is now available to the public, and provides an excellent foundation for understanding microbial diversity and function in extreme thermophilic habitats, as well as potential applications in bio-energy and bio-processing.

Understanding the Ecological and Physiological Roles of the *Verrucomicrobium* sp. Strain TAV2 in the Termite Hindgut through a Systems Biology Approach

Jantiya Isanapong,[1] Austin G. Willis,[1] Patrick Chain,[2] Stephen Callister,[3] and Ljiljana Pasa-Tolić,[3] Thomas M. Schmidt,[4] and **Jorge L.M. Rodrigues**[1]* (jorge@uta.edu)

[1]University of Texas, Arlington; [2]Los Alamos National Laboratory, Los Alamos, New Mexico; [3]Pacific Northwest National Laboratory, Richland, Washington; and [4]Michigan State University, East Lansing

Wood-feeding termites harbor an entire microbial community orchestrated to transform cellulose and hemicellulose into oligosaccharides, H_2, and CH_4. In order to have a better understanding of genetic capabilities of its members, we have sequenced the genome of the termite-associated *Verrucomicrobium* strain TAV2. Although at a draft stage, the TAV2 genome has revealed important genetic attributes for overall cellulose degradation. Firstly, the genome contains approximately 187 genes related to carbohydrate utilization and among those, genes encoding for glycosyl hydrolases (family 5 of cellulases), xylanase and acetyl xylan esterase, 1,4-β-glucanase, and α-L-arabinofuranosidase. These are potential sources of catalysts for funneling lignocellulose to oligosaccharides. Secondly, genes encoding for nitrogen fixation are present, suggesting a possible role of TAV2 on the maintenance of N stocks in the high C/N ratio termite hindgut environment. Thirdly, a *cbb*$_3$-type cytochrome oxidase encoding gene was identified, providing sub-oxic conditions in the surrounding environment. We hypothesize that TAV2 is important for controlled consumption of O_2, maintaining an anoxic hindgut core. Physiological experiments confirmed that TAV2 has a higher growth rate at 2% O_2 in comparison to cells grown under 20% O_2. We have developed an oligonucleotide microarray containing 4,022 coding sequences and performed a competitive hybridization with mRNA extracted from TAV2 cells grown under the those two O_2 concentrations. Comparisons between these two conditions revealed that 122 genes have a two-fold change in expression. Proteomic analysis of these samples uncovered 155 proteins uniquely present in cells cultured at 2% O_2, while 70 proteins were identified as unique to cells cultured at 20% O_2. Interestingly, many of these proteins have been previously classified as hypothetical proteins, indicating our limited knowledge about microbial life under hypoxic conditions.

Genome sequencing was supported by DOE/JGI (CSP186). A portion of this research was performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

PyroTagger: A Fast, Accurate Pipeline for Analysis of rRNA Amplicon Pyrosequence Data

Victor Kunin* (VKunin@lbl.gov) and Philip Hugenholtz

DOE Joint Genome Institute, Walnut Creek, California

Pyrosequencing of small subunit ribosomal RNA amplicons (pyrotags) is rapidly gaining popularity as the method of choice for profiling microbial communities because it provides deep coverage with low cost. However, the large amount of data, and errors associated with the sequencing technology present significant analytical challenges. Here we describe PyroTagger, a computational pipeline for pyrotag analysis. The pipeline consists of read quality filtering and length trimming, dereplication, clustering at 97% sequence identity, classification and dataset partitioning based on barcodes. To speed up the rate-limiting clustering step we developed a novel purpose-built algorithm called pyroclust. PyroTagger is highly scalable, capable of processing hundreds of thousands of reads within minutes on a single CPU. Version 1.0 is available for public use at http://pyrotagger.jgi-psf.org/.

The Open Journal – Social Network, Journal Club and Peer-Reviewed Journal with Automated Editors and Production

Victor Kunin* (VKunin@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

http://www.theopenjournal.org/

The Open Journal is designed to let science benefit from the current communication tools. The social network hosts profiles of scientists including education, research interests, work history and publications. Journal club allows exchanging opinions on publications. The peer-reviewed journal lets authors upload papers and have a complete control over the peer review process. Authors choose potential referees and the system automatically verifies that referees have appropriate qualifications and no conflicts of interests. The reviewer identities are public, ensuring both acknowledgment and accountability for referees. The production process is fully automated, rendering journal-formatted articles in HTML and pdf formats directly from authors' submission. Beyond accountability, transparency and open access, the system is designed for speedy publication process at low cost.

The Pan-Genome of *Emiliania huxleyi*

Alan Kuo[1]* (AKuo@lbl.gov), Betsy Read,[2] and Igor Grigoriev[1]

[1]DOE Joint Genome Institute, Walnut Creek, California and [2]California State University, San Marcos

De novo Genome Sequencing and Genome Closing of the Toxic Cyanobacterium *Planktothrix agardhii*

Rainer Kurmayer* (rainer.kurmayer@oeaw.ac.at) and Guntram Christiansen

Austrian Academy of Sciences, Institute for Limnology, Mondseestrasse, Mondsee

Genome sequencing clearly has a high potential to elucidate the physiological and ecological differentiation of toxin-producing and non-toxic strains in cyanobacteria. Transposable elements are typically involved in DNA rearrangements. Thus they are believed to form a major factor in the generation of genetic variation within populations of prokaryotes in general. Previously we described different insertion elements regulating the production of the hepatotoxin microcystin in the cyanobacterium *Planktothrix agardhii*. In addition gene clusters encoding the synthesis of toxic/bioactive metabolites may be flanked by transposable elements implying a potential involvement in the transfer of those gene functions. On the other hand it is well known that due to their repetitive nature transposable elements or insertion elements often hinder the automated assembling of the sequenced DNA molecules. Consequently, insertion elements in genomes of cyanobacteria typically have been only partially elucidated, as this part of genome information has not been completed.

The *de novo* genome sequencing of the microcystin-producing cyanobacterium *Planktothrix agardhii* (strain CYA126/8) by the 454 pyrosequencing technology (GS20, 23× coverage) resulted in 805 contiguous DNA sequences (4.97 Mbp genome size). Using paired end sequencing these 805 contiguous DNA sequences (contigs) were reduced to 23 contigs containing 370 gaps that were closed by PCR. However only by the construction of

a genome library (cosmid) and subsequent cosmid walking the 23 contigs could be further reduced finally resulting in three large contigs comprising 4.5 Mbp. Notably three contigs turned out to represent five plasmid sequences (5, 6, 50, 79, 115 kbp) which were erroneously assembled by the Newbler Assembly software. The reasons for this incorrect assembly will be presented.

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

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

[1]Los Alamos National Laboratory, Los Alamos, New Mexico; [2]University of Michigan, Ann Arbor; [3]Duke University, Durham, North Carolina; [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Washington State University, Pullman; [6]Smithsonian Environmental Research Center, Edgewater, Maryland; [7]Northern Arizona University, Flagstaff; [8]Western Illinois University, Malcomb; [9]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California

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

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

Using targeted (rDNA and functional genes) metagenomics, shotgun metagenomics, and quantitative PCR approaches, we are studying the impacts of over 10 yrs of elevated CO_2 on the soil microbial communities across these six different ecosystems. It is clear from prior research at these sites that soil microflora have responded to the climate change factors and that the populations and mechanisms underlying those responses are likely to be complex. Soil microbial communities might respond to elevated CO_2 in several different ways. For example, total biomass or biomass of major microbial groups may change. Certain populations may grow or be inhibited, resulting in measurable changes in community structure and composition. Alternately, the community may respond via alterations in metabolic activity without measurable changes in structure. Assessment of the bacterial communities in soils across the sites has shown that the bacterial communities have responded to elevated CO_2 in some ecosystems but not in others, and that the nature of the response is specific to that ecosystem. Responses have included changes in relative abundance of soil bacteria, as well as changes in community richness and composition. Where multiple factors were included in the field site experiment (e.g. soil depth, plant

species, ozone treatment), those factors were also found to alter soil bacterial biomass and composition. A similar assessment of the fungal communities has identified very different fungal populations that dominate the different ecosystems. Changes in relative abundance of fungal biomass and composition are correlated with elevated CO_2 at some of the sites. In order to facilitate fungal community comparisons, we have developed a new naïve Bayesian classifier to bin fungal LSU sequences into reliable, validated taxonomic units, and are establishing a large sequence training set for fungal LSU and ITS sequences.

Microbial Annotation at the Oak Ridge National Laboratory

Miriam Land* (landml@ornl.gov), Loren Hauser, Frank Larimer, Janet Chang, Cynthia Jeffries, Doug Hyatt, Chongle Pan, Tom Brettin, Daniel Quest, and Robert Cottingham

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

The Oak Ridge National Laboratory provides genome annotation for the JGI microbial genomes. In addition to annotating genomes, the ORNL effort also constructs tools and infrastructure to assist the scientific research community in the analysis and biological understanding of microbial genome sequences. In the currently funded effort, the specific aim is to apply and enhance gene recognition and modeling systems that support the phylogenetic scope of microbial genomes and communities being sequenced. High-throughput genome annotation and automation is used whenever possible to support the volume of sequences produced by the JGI.

ORNL provides automated annotation for draft and/or finished assemblies of archaeal and bacterial genomes sequenced by the JGI. This is made possible through the development and maintenance of robust scalable systems that accommodate the growth of genomic sequence data. The details of the final product and its delivery vehicle may be modified as changes are made to the mixture of sequencing technologies and program needs. Currently, the data are placed online in a web environment for easy access during the data embargo and after.

Annotation may include the prediction of genes, RNAs, repeats, operons, and other regions of interest in the DNA of genomes. The minimum set is protein-coding genes, tRNAs, and rRNAs. Protein coding genes will be identified using the locally developed tool Prodigal. The conceptual translations of predicted gene models are used to generate similarity search results and protein family relationships; from these results a metabolic framework is constructed and functional roles assigned. Repeats, tRNA genes and other structural RNA genes are identified with existing tools whenever appropriate. New tools are developed if quality, sensitivity, or accuracy can be improved by doing so. Annotation is made available to users through web page details and summaries, GenBank formatted files ready for submission, and IMG-ready files which can be downloaded.

ORNL stays flexible enough to adapt to the changing needs of the JGI microbial program. This may include 1) changes to the definition of minimal annotation, 2) annotation of metagenomes, 3) continued integration with the IMG web site, 4) annotation of transcriptomics, or 5) others not known at this time.

Comprehensive representation of microbial genomes requires deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, repressor and activator binding site calling, transcription terminators, and other functional elements. Sensor development is continuing to enhance access to these features. Linkage and integration of the gene/protein/function catalog to

phylogenomic, structural, proteomic, transcriptional, and metabolic profiles are being considered for development. The expanding set of microbial genomes comprises an extensive resource for comparative genomics: new tools can be developed for rapid exploration of gene and operon phylogeny, regulatory networking, and functional proteomics. ORNL will continue to seek opportunities to develop tools which provide users with better analysis and understanding of their genome.

Exploring the Aspen-Laccaria Mycorrhizal Interactome using NGS

Peter E. Larsen[1]* (plarsen@anl.gov), Geetika Trivedi,[2] Avinash Sreedasyam,[2] Vincent Lu,[1] Gopi K. Podila,[2] and Frank R. Collart[1]

[1]Biosciences Division, Argonne National Laboratory, Lemont, Illinois and [2]Department of Biological Sciences, University of Alabama, Huntsville

Many tree species that dominate forest ecosystems develop mutualistic symbiotic associations, so-called ectomycorrhizae, with soil fungi. The fungi contribute phosphorous, nitrogen and mobilized nutrients from organic matter and in return the fungus obtains up to 25% of the plant-derived carbohydrates. Knowledge of the molecular events associated with the development of the mycorrhizal system would facilitate our understanding of natural biological processes related to carbon sequestration, sustainability, and bioenergy. We used Solexa NGS and the JGI database of Poplar and Laccaria gene models and annotations to profile changes in gene expression associated with the transition to the mycorrhizal state. Free-living, early mycorrhizal interaction, and fully formed mycorrhizae conditions were considered, collecting a 134 million sequence reads from Laccaria, 151 million sequence reads from Aspen, and 25 million sequence reads for fully-formed mycorrhizae from mixed samples of Aspen and Laccaria for a total of 310 million sequence reads. This transcriptomic data was used to investigate key portions of the Aspen-Laccaria mycorrhizal interactome. Time course data for early mycorrhizal interactions was used to predict protein-protein interactions. Gene models differentially expressed between free-living and fully-formed mycorrhizae were mapped to KEGG pathways and used to construct a mycorrhizal metabolic network. A regulatory network was constructed by considering all expressed gene models in mycorrhizae, identifying those regulatory elements that were enriched for statistically significant differentially expressed gene models. These three proposed networks together comprise a complete representation of the Aspen-Laccaria mycorrhizal interactome and these predicted interactions are immediately useful for deriving hypothesis-driven, biological experiments.

Keywords: Ectomycorrhizae, next generation sequencing, systems biology, carbon metabolism, *Laccaria bicolor*, *Populus tremuloides*, symbiosis

Overcoming Some of the Challenges to Single Cell Genomics

Janey Lee* (JLee2@lbl.gov), Damon Tighe, Mei Wang, Stephanie Malfatti, Erika Lindquist, Feng Chen, Jan-Fang Cheng, and Tanja Woyke

DOE Joint Genome Institute, Walnut Creek, California

Single cell genomics, the amplification and sequencing of genomes from single cells, can provide a glimpse into the genetic make-up and thus life style of the vast majority of uncultured microbial cells, making it an immensely powerful and increasingly popular tool. This is accomplished by use of multiple displacement amplification (MDA), which

can generate billions of copies of a single bacterial genome producing microgram-range DNA required for shotgun sequencing. Here, we would like to address several challenges inherent in such a sensitive method and propose solutions for the improved recovery of single cell genomes. While DNA-free reagents for the amplification of a single cell genome are a prerequisite for successful single cell sequencing and analysis, DNA contamination has been detected in various reagents, which poses a considerable challenge. Our study demonstrates the effect of UV radiation in efficient elimination of exogenous contaminant DNA found in MDA reagents, while maintaining Phi29 activity. Second, MDA is subject to amplification bias, resulting in uneven and sometimes insufficient sequence coverage across the genome. In a post-amplification method, we employed a normalization step within 454 Titanium library construction in which populations of highly abundant sequences were specifically targeted and degraded from the library via duplex-specific nuclease, resulting in decreased variability in genome coverage. While additional challenges in single cell genomics remain to be resolved, the two proposed methodologies are relatively quick and simple and we believe that their application will be of high value for future single cell sequencing projects.

The New Science of Metagenomics: Bioprospecting the Secrets of Microbial Communities

Luen-Luen Li[1,4]* (luenlee@bnl.gov), Sean M. McCorkle,[1,4] Yian-Biao Zhang,[1] Susannah G. Tringe,[2] Tanja Woyke,[2] William S. Adney,[3,4] Shi-You Ding,[3,4] Michael Himmel,[3,4] Safiyh Taghavi,[1,4] and Daniel van der Lelie[1,4]*

[1]Brookhaven National Laboratory, Upton, New York; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]National Renewable Energy Laboratory, Golden, Colorado; and [4]BioEnergy Science Center, U.S. Department of Energy

In the environment, microorganisms have evolved and accumulated remarkable physiological and functional diversity, and constitute the major reserve for genetic diversity on earth. Using metagenomics, this genetic diversity can be accessed without the need of cell cultivation. Microbial communities and their metagenomes, isolated from biotopes with high turnover rates of recalcitrant lignocellulosic plant cell wall biomass, have become a major resource for the bioprospecting and discovery of new biocatalytics (enzymes) for various industrial processes, including the production of biofuels from plant feedstocks.

Our work aims to understand the diversity and metabolic capabilities of an anaerobic microbial community actively decaying poplar biomass. Metagenomic DNA was isolated and sequenced using 454-GS-FLX Titanium pyrosequencing. Approximately 720Mbp reads were generated which assembled into 198,375 contigs with a total size of 128 Mb, on which 653,488 putative genes were identified. 16S/18S rRNA libraries and 454-pyrotag sequencing, dinucleotide frequency analysis with Aglomerative clustering (AGNES), and ensemble's G/C content analysis all suggested that the community is dominated by 5 demarcated phylogentic groups: Two Bacteroides groups, Firmicutes, Magnetospirillum and previously uncultured Firmicutes. Links between phylogentic groups and functions (COG/Pfam assignments) are currently under investigation.

Approximately 4,000 glycosyl hydrolase (GHase) homologues were identified using blastx searches against CAZy database. A comparison of GHase composition between our decaying poplar biomass metagenome and the termite gut metagenome is in process. Based on homology to GHase families/activities of interest (key enzymes for efficient decay of plant cell wall recalcitrants) and quality of sequences, candidates were selected

for further investigation. One of the bottlenecks, however, is that many putative genes are incomplete or incorrectly assembled, making their full length cloning a time consuming and costly approach. Despite this limitation, full-length open reading frames of various GHases were obtained using inverse PCR and DNA walking, and subsequently cloned into an expression vector for expressing in *E. coli*. Protein purification and characterization are presently in process.

16S rRNA Sequence Taxonomy Assignment by Repetitive Oligonucleotide

Kuan-Liang Liu[2]* (kliu@lanl.gov), Gary Xie,[1] Patrick S. Chain,[1]* Cheryl R. Kuske,[1] and Nicolas W. Hengartner[2]

[1]Bioscience Division, Genome Science and [2]Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico

Massively parallel pyrosequencing of hypervariable regions from small subunit ribosomal RNA (SSU rRNA) genes can provide rapid, inexpensive analysis of microbial community composition. In order to get deeply understanding of the community from these short reads, it is very important to assign taxonomy information to them. To that end, the Ribosomal Database Project developed the 'Classifier' to characterize prokaryote communities sampled using ribosomal DNA markers (rDNA). The classifier first extracts overlapping 8-mers from the read as features, then assign the read to the genus that has the most abundant feature frequencies. It works extremely fast and can get accurate assignment even with illumina reads extracted from adequate primers. However, the model did not consider the repetitive 8-mers in each reference sequence. Since rRNAs typically have many secondary structures due to reverse complement sequence, we believe that these additional appearances of 8-mers will provide more information during taxonomy assignment. Here, we propose two modifications by utilizing repetitive 8-mers and compare the performance with the original RDP classifier.

Keywords: pyrosequencing, taxonomy assignment, 16S rRNA, hypervariable regions

Protein Kinases as a Focused Model for Evolutionary Genomics and Inferring Biology from Genome Sequence

Gerard Manning* (manning@salk.edu), Mike Dacre, Yufeng Zhai, Eric Scheeff, Aaron Legler, Ana Rodrigues, Anna Luan, and Olivia Gardiner

Salk Institute for Biological Studies, La Jolla, California

We study the evolution of protein kinases to provide a focused, curated view of gene evolution, and to better understand cell signaling and control mechanisms. The kinome consists of ~2% of genes in most eukaryotes, and these collectively control the vast majority of cellular functions. We infer about 50 distinct kinase types in the last common eukaryotic ancestor, expanding to over 500 kinases in human. This provides a model that covers a large but tractable number of genes, with many distinctive and conserved subsets.

Using HMM profiles, intron positioning and curation, we can substantially improve detection and sequence accuracy of kinase genes across dozens of eukaryotic species, and greatly enhance orthology predictions. Kinomes from basal metazoans and protist relatives show early development of many kinases classes that were subsequently lost in insects and

nematodes, and reveal large independent expansions of individual classes in most genomes.

Prediction of kinases across 28 vertebrate genomes provides tools to improve gene prediction, by use of conserved intron positions, knowledge of variable rates of evolution, and better evaluation of genomic assembly problems. The combined vertebrate kinome allows us to map the evolutionary constraints on every residue of each kinase, and to prioritize SNPs and somatic mutations for functional impact, as well as to detect functional shift in specific lineages. A similar approach can be applied to other clusters of related genome, including plants and fungi.

We are applying lessons from the kinase work to whole proteome orthologies, using 7 nematode genomes as a test case. Standard tools such as OrthoMCL and InParanoid cluster only a minority of nematode genes into singular orthology groups. We have substantially improved their predictive power with multi-level clustering, code tweaks and the integration of synteny and whole-genome alignments to both maximize the discovery of ancestral genes and to follow the evolutionary dynamics from that point to current genomes. We are also using multiple approaches to build orthologies between eukaryotic phyla, providing much larger coverage of genomes than current methods.

Immortal DNA—The Chief Architect of the Nature's Biological Phenomena with "Gene Re-Cycling" through Flora and Fauna

Zachariah Mathew* (nammal6@aol.com)

Thajema Scientific Book Publishers, West Orange, New Jersey

So far the modern science has revealed to us that by way of ecological re-cycling system, the plants take up and re-cycle the nutrients from the dead flora and fauna. However, it is not yet fully understood whether the plant roots can take not only the nutrients but also the genes of the dead flora and fauna.

If the roots of the plants are taking the genes of the dead flora and fauna, along with the other organic and inorganic nutrients, then the plant and its parts will also contain the foreign genes of those flora and fauna. The plant and its fruits containing such foreign human dominant genes when consumed by other people in the locality or town, the genes may enter their body through horizontal gene transfer (HGT). The expression of the dominant genes will eventually take place in the next generation of people in the locality, giving some of the dominant genetic characters in the next generation, producing the ethnicity. Thus one of the Nature's Biological Phenomena like human ethnicity can be attributed to Gene Re-cycling hypothesis.

Gene Re-Cycling Hypothesis: By corroborating the Nature's biological phenomena, (such as Ethnicity, Re- appearance of extinct animals, Cyclic occurrence of viral diseases, Infectious disease outbreaks, by eating the fresh vegetables and fruits grown in soil contaminated with pathogenic organisms, etc.) and current scientific literature on transgenic plants, transgenic animals, role of gene regulatory proteins, viral encoded regulatory proteins, bacterial regulatory proteins, cell senescence, human genomic studies, DNA sequence and re-sequence studies etc. it is hypothesized that genes of every flora and fauna are recycled from generation to generation.

Experiment on the Role of Naturally Developed Transgenic Plants and Animals: in Gene Recycling: Experiment is planned to develop a natural transgenic paddy plant

without gene manipulation. Gene re-sequencing of the paddy seed will be conducted to confirm the transgenic character of the paddy seed. The naturally developed transgenic paddy seeds will then be fed to male and female mice. At maturity mice gene will be resequenced to note the presence of transgenes from paddy seeds. If found the male and female mice will be mated and the offspring from those transgenic mice will also be resequenced to confirm the Gene Re-cycling hypothesis.

Project Relevance to Department of Energy (DOE) mission: Societal, Economic and Scientific Importance: A scientific proof of "Gene-Recycling" hypothesis will enable the scientific world to understand the hidden truth behind the Nature's biological phenomena such as Cyclic occurrence of viral diseases, Re-appearance of extinct animals, Public health importance of outbreak of infectious diseases by eating fresh vegetables / fruits grown in soil infected with pathogenic organisms and the root cause of ethnicity.

Scientific Merit: Proving "Gene Recycling" hypothesis by scientific experiments such as Community Sequencing Program and genomic studies will change the present concept of mortality of living organisms (cells) to immortality.

Gene Prediction and Size Reduction on Metagenomic Datasets

K. Mavrommatis* (KMavrommatis@lbl.gov), A. Pati, N. Ivanova, M. Hunteman, P. Williams, and N.C. Kyrpides

DOE Joint Genome Institute, Genome Biology Program, Walnut Creek, California

Metagenomics allows the study of complex microbial communities. In every metagenomic analysis the major goal is the prediction of genes and the identification of their function and the metabolic reconstruction of the community. Ab initio gene calling on short sequences is not accurate due to the short size of the sequences, and the presence of sequencing errors. Similarity based tools don't have these problems, however their application is limited to the detection of protein coding genes similar to whatever is already known, and are computationally intensive. Furthermore, new sequencing technologies and in particular pyrosequencing (454) typically result in datasets with high numbers of sequences to be analysed, a difficult task for any researcher.

Here we report a comparative analysis of the performance of three gene calling methods, widely used for the prediction of genes on metagenomic dataset, by means of sets of simulated datasets. We also present a protein assembly based method which allows the "compression" of the metagenomic datasets in order to allow a more efficient storage and handling.

Functional Annotation of *Fibrobacter succinogenes* Carbohydrate Active Enzymes

David Mead[1]* (dmead@lucigen.com), Julie Boyum,[1] Colleen Drinkwater,[1] Krishne Gowda,[1] David Stevenson,[2] Paul Weimer,[2] and Phil Brumm[3]

[1]Lucigen and Great Lakes Bioenergy Research Center, Middleton, Wisconsin; [2]USDA-ARS, Madison, Wisconsin; and [3]C5-6 Technologies and Great Lakes Bioenergy Research Center, Middleton, Wisconsin

Fibrobacter succinogenes is a predominant cellulolytic bacterium that degrades plant cell wall biomass in ruminant animals, and is among the most rapidly fibrolytic of the known

mesophilic bacteria. A dozen cellulolytic enzymes have been expressed and characterized previously, and an outdated partial genome sequence indicates that there are at least 33 unique glycosyl hydrolases encoded by *F. succinogenes*. In order to better understand plant cell wall degradation we have developed new tools to capture, express and identify many of the carbohydrate active enzymes (CAZYmes) from this microbe. The complete genome sequence of Fsu was finished by the DOE Joint Genome Institute in late 2009, contributing to the growing database of cellulolytic microbes. Preliminary analysis indicates that *F. succinogenes* contains ~104 glycosyl hydrolase and 63 CBM-containing genes.

Based on the genomic sequencing results, the number of *F. succinogenes* genes annotated as CAZymes far exceed those that have been experimentally determined by conventional enzymatic approaches. One of the goals of this work is to functionally characterize all the putative glycosyl hydrolase genes from Fsu, as bioinformatic analysis is an inadequate proxy for actual activity results. Before the genome sequence was available we developed a robust method to enzymatically capture functionally active CAZymes in *E. coli*. Using new expression tools developed at Lucigen and C5-6 Technologies and a multi-substrate screen for β-xylosidase, xylanase, β-glucosidase and endocellulase activities, we generated and screened 5760 random shotgun expression clones for these activities. This represents ~2 X genome expression coverage. 169 positive hits were recorded and 33 were unambiguously identified by sequence analysis of the inserts. Eliminating duplicates, 24 unique CAZyme genes were found by functional screening, or 40% of the ~60 genes present in this genome potentially detectable by the multiplex assay. Several previously uncharacterized enzymes were discovered using this approach. With the full genome sequence available we will attempt to express and characterize all of the recognizable CAZymes, as well as the CBM-containing genes for actual enzyme activity. The active enzymes will also be sent to other partners in the GLBRC to assess their ability to deconstruct plant biomass.

Genomic Analysis of Microbial Communities in Chronic Wounds

Johan H. Melendez[1]* (jmelend3@jhmi.edu), Lance B. Price,[2] Yelena M. Frankel,[1] Anne Han,[1] Emmanuel Mongodin,[3] Gerald S. Lazarus,[2] and Jonathan M. Zenilman[1]

[1]Johns Hopkins Medical Institutions, Baltimore, Maryland; [2]Translational Genomics Research Institute, Flagstaff, Arizona; and [3]University of Maryland, Baltimore

The microbial communities residing in chronic ulcers may play a crucial role in the persistence of chronic wounds. However, these communities have not been properly described due to the limitations of traditional culture-based approaches. In order to better describe the bacterial ecology of these wounds, we have analyzed tissue samples from 36 chronic wound patients using 16S rRNA gene pyrosequencing and culture. The 16S-based analyses revealed that the communities of organisms inhabiting these wounds are more than diverse than previously thought and approximately 4 times greater than estimated by culture-based analyses. Large and diverse populations of fastidious anaerobes were exclusively identified by the genomic analyses further highlighting the limitations of traditional culture methods. The wound microbial ecology of patients treated with antibiotics was different and less diverse than from untreated patients. Taken together, these results suggest that the microflora of chronic wounds is more diverse than previously described, that anaerobes are common, and that antibiotic use results in decrease microbial diversity. 16S rRNA based pyrosequencing is an excellent tool for the characterization of chronic wound microbiota and elucidating the microbial ecology of these wounds can help guide appropriate antimicrobial therapy.

A Metagenomic Exploration of Lignocellulose Degradation in the Bovine Rumen

C.D. Moon* (christina.moon@agresearch.co.nz), D. Gagic, D. Li, C. Sang, E. Altermann, W.J. Kelly, S.C. Leahy, and G.T. Attwood

Food, Metabolism and Microbiology Section, AgResearch Grasslands, Palmerston North, New Zealand

Significant advances are required to enable the cost-effective conversion of lignocellulosic biomass for the production of biofuels and biomaterials. Therefore, understanding the diversity of biological mechanisms that mediate lignocellulose degradation is of considerable interest. The rumen is the fermentative forestomach of ruminant animals, and the resident microbiota has evolved to become arguably one of the most effective microbiomes specialised in biomass degradation. Rumen microbial communities are taxonomically diverse, and numerically dense, and they interact to effectively and rapidly convert the ingested forage plant material into substrates which are readily metabolised by the host. The microbes that are attached to, or, closely associated with, the digesta are thought to encode the functional "core" of biomass conversion processes in the rumen, and are likely to express much of the carbohydrate-active enzymes necessary to hydrolyze the complex polysaccharides that comprise plant cell walls. However, as the majority of these microbes are unable to be cultured, and community composition can vary markedly depending on diet, the hydrolytic and metabolic potential of these microbes is still largely undefined. This study aims to provide detailed insight into the variety of mechanisms employed by bovine rumen microbiota to degrade forage, by characterising the enzymes and enzyme consortia from the microbes that colonise the digesta *via* functional and sequence-based metagenomic approaches at a depth of coverage significantly greater than performed previously. Metagenomics allows the direct study of environmental genetic material, enabling access to entire microbial communities. We have undertaken a functional metagenomics approach (which has the potential to identify highly novel activities) to investigate the core processes of ruminal biomass degradation. The rumen contents of two pasture-grazed dairy cows were fractionated to obtain the plant-associated and plant-adherent microbiota. High molecular weight metagenomic DNA was isolated from each fraction and used to construct large-insert fosmid libraries in pCC2FOS. Over 11,000 library clones were screened through agar plate based bioassays for various glycosyl hydrolase and carbohydrate esterase activities involved in the degradation of lignocellulose and starch. A total of 386 clones (3.42%) exhibited activities in one or more of the bioassays at the following frequencies: endoglucanase (0.35% of clones), cellobiohydrolase (0.48%), beta glucosidase (0.65%), endoxylanase (0.38%), xylosidase (0.65%), arabinofuranosidase (0.70%), phenolic acid esterase (0.23%), and amylase (0.61%). Twenty bioactive clones expressing multiple, or potentially high activities, were analysed by transposon mutagenesis, revealing novel genes with predicted fibre-degrading activities, putative regulators of these genes, and genes of unknown function. In collaboration with the JGI, remaining bioactive fosmids will be fully sequenced and annotated to identify genes responsible for bioactivities. Furthermore, rumen microbial community structure will be investigated by analyses of 16S rRNA gene sequences. We plan to further characterize novel bioactive genes and their products to expand our knowledge of the mechanisms involved in ruminal fibre degradation. These may lead to important applications in biofuel production and agricultural systems.

Conservation of Regulatory Networks in the *Phytophthora sojae* Genome

Paul F. Morris[1]* (pmorris@bgsu.edu), Trudy Torto-Alibo,[2] Sucheta Tripathy,[2] Felipe Arredondo,[2] Nicole Vanduzen,[1] Alexandra Schmucker,[1] and Brett Tyler[2]

[1]Biological Sciences, Bowling Green State University, Bowling Green, Ohio and [2]Virginia Bioinformatics, Virginia Tech, Blacksburg

Computational predictions of protein-protein interactions have been used to predict associations that are conserved across the major eukaryotic lineages of plants, animals and fungi. These predictions rely on identifying orthologous pairs of proteins that have been experimentally verified to predict protein-protein interactions in other organisms. Comparisons across these kingdoms may be robust for many processes because there has been very little horizontal exchange of domains after separation of these Kingdoms. In contrast, analysis of diatom and oomycete genomes have revealed multiple examples of horizontal transfer events in these genomes. These include both transfer of genes to the host nucleus from the photosynthetic endosymbiont in the common ancestor of oomycetes and diatoms, along with the acquisition of bacterial genes. In spite of these inherent problems, a predicted interactome is a valuable first step to help map out conserved signaling networks in oomycete genomes. We first used the ortholog search strategy of reciprocal smallest distance to identify conserved orthologs in *Phytophthora sojae* with respect to the *A. thaliana*, human and S. *cerevisiae* genomes. Conserved orthologs were then used to identify ortholog pairs in the predicted *A. thaliana* and *S. cerevisae* interactomes. This strategy captured 14,235 of 72,266 interactions in the *A. thaliana* interactome and 13,298 of 53,301 interactions from the *S. cerevisiae* interactome. Expression data from 9039 genes in 42 microarray experiments of germinating zoospores and mycelia included expression data on 57% of the orthologs in the *A. thaliana* human and *S. cerevisiae* genomes. Coexpression analysis of P sojae genes with orthologs to only the *A. thaliana* genome, revealed more than 9300 predicted interactions with a Pearson correlation coefficient greater than R=0.7. However only 46 ortholog pairs were also part of the set of 14,235 interactions captured from the *Arabidopsis* interactome. While genes can be components of regulatory networks and not have high levels of co-expression, the low number of co-expressed genes that are also part of the predicted interactome suggest that no unique regulatory networks from the ancestral endosymbiont genome have been conserved in oomycetes. Moreover, only 25 % of interaction pairs from the predicted interactome had R scores greater than 0.5. Thus even conserved regulatory networks may be regulated different in oomycetes from other eukaryotes.

Ecological and Evolutionary Relevance of the Genomes of Thermophilic Fungi

Donald O. Natvig[1]* (dnatvig@unm.edu), Joslyn Bustamante,[1] Eric Ackerman,[2] Bryce Ricken,[2] Randy Berka,[3] Adrian Tsang,[4] and Amy J. Powell[1,2]

[1]Department of Biology, University of New Mexico, Albuquerque; [2]Sandia National Laboratories, Albuquerque, New Mexico; [3]Novozymes, Inc., Davis, California; [4]Center for Functional and Structural Genomics, and Department of Biology, Concordia University, Montreal, Canada

Although thermophilic fungi have received substantial attention for potential use in industrial applications, important questions remain regarding their distributions, ecology, life cycles and diversity. If these organisms are to be exploited fully in industry, including the production of biofuels, it will be important to better characterize aspects of their

fundamental biology. We are examining the ecology and evolution of major groups of thermophilic fungi, combining information from new genome sequences with field and laboratory studies of isolates obtained from arid ecosystems. Within the Ascomycota, thermophily has arisen independently in at least two groups, the Chaetomiaceae and the Eurotiales. The genomes of three closely-related members of the Chaetiomiaceae have been sequenced, and at least eight well-curated genomes are available for mesophilic and thermolerant taxa within the Eurotiales. Two sequenced members of the Chaetomiaceae, *Thielavia terrestris* and *Sporotrichum thermophile*, are thermophiles, while the third, *Chaetomium globosum*, is not. Analysis of these genomes underscores key unresolved issues within the family. One is that taxonomic and nomenclatural problems are rampant among thermophilic species. For example, in many cases a single genus name has been applied to species from distantly related groups. This includes *Sporotrichum*, a name that has been used for diverse filamentous fungi and is perhaps properly applied to particular anamorphic species in the Basidiomycota. Moreover, it would appear that the genus *Chaetomium* as currently viewed is paraphyletic. Aside from engendering confusion about naming, the poor understanding of phylogenetic relationships within the Chaetomiaceae makes it impossible to evaluate where thermophily has been gained or lost in the family, which in turn hampers efforts to probe the molecular basis of thermophily. Our field studies of litter, biological crusts, rhizosphere soils and herbivore dung from the Sevilleta Long-Term Ecological Research Site in central New Mexico reveal a large diversity of thermophilic fungi from both the Chaetomiaceae and the Eurotiales. One early conclusion based on analyses of ribosomal ITS and other sequences is that species diversity in both major groups of thermophilic fungi has been underestimated.

Meeting Data-Intensive Computing Challenges in Genomics

Christopher Oehmen, Ian Gorton, and **Lee Ann McCue*** (leeann.mccue@pnl.gov)

Computer Science and Mathematics Division, Pacific Northwest National Laboratory, Richland, Washington

Constantly evolving high-throughput sequencing technologies are driving an exponential increase in genomic data, and generating an analysis bottleneck. To keep pace with this data explosion and relieve the bottleneck, we are developing efficient parallel implementations for biological data analysis tasks, such as pairwise alignment and homology inference. For example, our high performance implementation of BLAST, ScalaBLAST, operates by distributing the work of the sequence analysis tasks across many processors with a dynamic, fault resilient task scheduling layer to manage the distribution of data files and work across multiprocessor systems. We have also developed a component-based integration platform, MeDICi, to facilitate the creation of complex analytical workflows handling large datasets and high volume data streams. Using high performance applications in this flexible framework we are eliminating the bottleneck of sequence alignment and homology detection from genomics analyses. The power of this approach is that 1) it can be changed to suit a broad array of lines of investigation, and 2) it allows hypotheses to be formulated from high-throughput data.

Genomics of the Ethanol-Producing *Zymomonas mobilis*

Katherine M. Pappas* (kmpappas@biol.uoa.gr)

Department of Genetics and Biotechnology, Faculty of Biology, University of Athens, Ilissia, Athens, Greece

Zymomonas mobilis is an ethanol-producing α-proteobacterium, long known for its natural involvement in plant-juice fermentations and currently considered a highly potent candidate for large scale bioethanol production. *Z. mobilis* is a suitable platform biocatalyst for several reasons: it surpasses yeasts in ethanol production rates and purity, has simpler fermentation requirements, is desirable for also fine-chemical production, is safe (GRAS), and it harbors a small ca. 2-Mb genome, which renders it more amenable to genetic manipulations compared to eukaryotes or even to other bacteria.

In order to understand the biology of *Z. mobilis,* six different strains belonging to two of its major subspecies (*mobilis* and *pomaceae*) and isolated from various parts of the globe, are under current analysis at JGI-DOE and the University of Athens (CSP_788284, 2008). The genome of strain ZM4 (ATCC 31821), the most robust ethanol-producer and a strain currently studied in U.S. academic and national laboratories, was sequenced in its entirety by sequencing its plasmid genome at JGI and by thus complementing its chromosomal genome already sequenced at Macrogen Inc. ZM4 was also reannotated,[1] and was recently chosen to receive additional transcriptome sequencing, which will improve gene prediction and reveal gene expression status at various conditions (CSP_52, 2010). The genome of NCIMB 11163, a British ale infecting strain, was also finished and proved to bear unique regions compared to ZM4 in both its chromosome and plasmids, among which a complete set of conjugal transfer genes that provides evidence for lateral gene exchange.[2] In finishing are currently the genomes of the type-strains of the two examined subspecies: that of ATCC 10988 (subsp. *mobilis*) originating in Mexico, and ATCC 29192 (subsp. *pomaceae*) originating in the U.K. ATCC 10988 is exceptional compared to the other strains in that it appears to have a most fluid genome, divided into no less than 9 replicons (a single chromosome and eight plasmid species). ATCC 29192 on the other hand, is the most deviant in total sequence homology, which complies well with its overall phenotypic difference and distinct classification. Lastly, strain CP4, a Brazilian isolate kin to ZM4 and mostly used in Canadian facilities, and ATCC 29191, a robust fermenter originating in Zaire, are also close to finishing. This on-going project aims to offer a sound basis for comparative and functional genomics analysis for this interesting ethanol-producing organism. It also aims to determine the fundamental core-, accessory-, and pan-genome complement for the *Z. mobilis* species, the knowledge of which will directly contribute to synthetic biology applications and designer strain built.

1. Yang, S, Pappas, KM, Hauser, LJ, Land, ML, Chen, G-L, Hurst, GB, Pan, C, Kouvelis, VN, Typas, MA, Pelletier, DA, Klingeman, DL, Chang, Y-J, Samatova, NF, and Brown, SD (2009) Improved genome annotation for *Zymomonas mobilis*. *Nature Biotechnol.* 27: 893.

2. Kouvelis, VN, Saunders, E, Brettin, TS, Bruce, D, Detter, C, Han, C, Typas, MA, and Pappas, KM (2009) Complete genome sequence of the ethanol producer *Zymomonas mobilis* NCIMB 11163. *J. Bacteriol.* 191: 7140.

Large Gap Size Paired-End Library Construction for Second Generation Sequencing

Ze Peng* (ZPeng@lbl.gov), Matthew Hamilton, Jeff Froula, Aren Ewing, Brian Foster, and Jan-Fang Cheng

DOE Joint Genome Institute, Walnut Creek, California

Fosmid or BAC end sequencing plays an important role in de novo assembly of large genomes like fungi and plants. However construction and Sanger sequencing of fosmid or BAC libraries are laborious and costly. The current 454 Paired-End (PE) Library and Illumina Jumping Library construction protocols are limited with the gap sizes of approximately 20 kb and 5 kb, respectively. In the attempt to understand the limitations of constructing PE libraries with greater than 30Kb gaps, we have purified 18, 28, 45, and 65Kb sheared DNA fragments from yeast and circularized the ends using the Cre-loxP approach described in the 454 PE Library protocol. With the increasing fragment sizes, we found a general trend of decreasing library quality in several areas. First, redundant reads and reads containing multiple loxP linkers increase when the average fragment size increases. Second, the contamination of short distance pairs (<10Kb) increases as the fragment size increases. Third, chimeric rate increases with the increasing fragment sizes. We have modified several steps to improve the quality of the long span PE libraries. The modification includes (1) the use of special PFGE program to reduce small fragment contamination; (2) the increase of DNA samples in the circularization step and prior to the PCR to reduce redundant reads; and (3) the decrease of fragment size in the double SPRI size selection to get a higher frequency of LoxP linker containing reads. With these modifications we have generated large gap size PE libraries with a much better quality.

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

NA2Dsearch: Fast and Easy Tool for Secondary Structure Searches Through Large Datasets in Parallel

Hlubuček Petr, **Mokrejš Martin*** (mmokrejs@iresite.org), and Pospíšek Martin

Faculty of Science, Charles University, Prague, Czech Republic

We have developed a search tool for RNA secondary structure motifs. The program searches for a user defined motif in a database of secondary structures (optionally restrained by primary sequence written in IUPAC codes).

The program aims to be usable for a wide constituency of users lacking extensive computer skills. Thus a comfortable graphical user interface (GUI) is introduced. A motif descriptor composed of single- (ss) and double-stranded (ds) regions can be constructed interactively through the GUI. Each ss- or ds-region is described by its minimal, maximal and optimal length (optionally also by primary sequence as already noted) and finally by a score weight expressing significance of a particular region. To facilitate more realistic queries the ds-regions can be detected even while being interrupted by a bulge(s) or even inner-loop(s).

A tree representation of secondary structure is used internally. Each ss- or ds-region is represented as a single node of the tree. The algorithm traverses the tree and uses non-deterministic finite automaton (NDFA) to find the motif in the database of structures.

The program can process large datasets – a search for a tRNA-like motif (without a pseudoknot) through 10 000 structures on average each 226bp long takes 9 to 36 seconds to our program depending on strictness of the query (either no bulges allowed at all, or only one bulge or one inner-loop, or up to three bulges and 3 inner-loops) on a quad-core, 2.6GHz CPU desktop computer. Benchmarks confirm nearly linear time complexity in respect to both count of structures in the target database(s) and their size. Memory requirements for the analysis are enforced only by a number of resulting hits that have to be held in memory before sorting by score as the database can optionally be read directly from a hard disk. NA2Dsearch extends family of RNA secondary structure search programs and offers a comfortable GUI, high speed and ability to deal with vast datasets. It is similar to the RNAmotif program but searches for a structural motif in a database of secondary structures (unlike RNAmotif that searches through primary sequences). It is implemented in Java as a parallel application and can be run in a batch mode.

The NA2Dsearch can be used to create search descriptors for RNAmotif and execute it as an alternative search engine through primary sequences. To further ease manual work NA2Dsearch can directly execute RNAshapes and perform a stuctural search through generated representative conformational shapes.

The program will be available through http://www.iresite.org/NA2Dsearch in July 2010. This work was supported by grants of the Ministry of Education, Youth and Sports of the Czech Republic (#LC06066) and Czech Science Foundation (#P305/10/J026).

Completion of the *Dekkera* (*Brettanomyces*) *bruxellensis* Genome Sequence

Trevor Phister,[1] Linda Bisson,[2] Jure Piskur,[3] Fred Dietrich,[4] Thomas Henick-Kling,[5] Scott Baker,[6] and Steven Gray[1]* (srgray@ncsu.edu)

[1]North Carolina State University, Raleigh; [2]University of California, Davis; [3]Lund University, Lund, Sweden; [4]Duke University, Durham, North Carolina; [5]Washington State University, Richland; and [6]DOE Pacific Northwest National Laboratory, Richland, Washington

*Dekkera bruxellen*sis (anomorph *Brettanomyces*) is a hemiascomycetes yeast with a genome size ranging from ~20 to ~30 Mb. It is a promising candidate for fuel ethanol production from both traditional starch-based feed stocks as well as lignocellulose based feed stocks. Its metabolism is similar to *Saccharomyces,* exhibiting anaerobic growth and Crabtree-positive fermentation. In one large-scale continuous fermentation, *D. bruxellensis* replaced *Saccharomyces* and the fermentation continued at levels of ethanol production equal to *Saccharomyces* fermentations for two years at pH 3.5 in the presence of a *Lactobacillu*s. This suggests that *Dekkera* be considered as a primary organism for fuel ethanol production. It has a high ethanol tolerance and does not generally grow until completion of the *Saccharomyces* fermentation as part of the microbial succession in most alcoholic fermentations, including fuel ethanol. In corn mash fermentations *Dekkera* growth rate depends on aeration, with higher aeration allowing *Dekkera* to negatively impact fermentations. While industrial aeration levels increase *Saccharomyces* ethanol production to anaerobic fermentations, they also allow *Dekkera* to affect ethanol production. Better understanding of *Dekkera* may allow novel control methods and design of more efficient, higher yielding fuel ethanol fermentations. *Dekkera* are also useful in

ethanol production from lignocellulose feedstocks because of resistance to hydroxycinnamic acids and phenolics, inhibitors formed during pretreatment, the ability to utilize sugars not used by *Saccharomyces* including cellobiose and arabinose, and its ability to perform simultaneous saccharification and fermentation (SSF), which produces more ethanol using less cellulase than other lignocellulose fermentations. Furthermore, *Dekkera*'s higher growth temperature allows for increased hydrolysis by cellulases during SSF as the fermentations can be conducted closer to their optimal temperatures. Additionally, *Dekkera* may prove useful in production of acetic acid because it produces more acetic acid compared to currently used organisms.

In this work, we will complete the *D. bruxellensis* CBS 2499 genome sequence. Over 40% of the genome is already sequenced with 1% of detected loci being polymorphic, a G+C content of 40.2%, and an estimated 7430 protein-coding ORFs. This sequence will enhance genetic engineering of *Dekkera* and other yeasts for the production of ethanol, acetic acid, and other industrially important chemicals. The *Dekkera* sequence is not only directly relevant to the DOE's goal of increasing ethanol production from biomass, but it will also allow researchers to address fundamental questions in evolution. The sequence of *D. bruxellensis* will provide more data for comparisons of parallel evolution with *Saccharomyces*. Both yeasts inhabit the same environments but have alternative metabolic strategies. It may be possible to exploit these differences in order to increase the metabolic capabilities of *Saccharomyces* by enhancing utilization of xylose and arabinose. One hurdle in the genetic engineering of *Saccharomyces* for use of xylose and other five carbon sugars is maintaining a positive redox balance. The initial *D. bruxellensis* sequencing found an alternative oxidase (*AOX1*), suggesting an additional pathway to help reoxidize NADH. Understanding this pathway in *D. bruxellensis* may allow for overcoming the redox balance difficulties.

Bioprospecting for Bacteria with Ligno-Cellulolytic Potential from Fungus-Growing Termites

Michael Poulsen[1,2]* (poulsen@bact.wisc.edu), Garret Suen,[1,2] Sandye Adams,[1,2] Duur K. Aanen,[3] Wilhelm de Beer,[4] Susannah G. Tringe,[5] Kerrie Barry,[5] Lynne A. Goodwin,[6] and Cameron R. Currie[1,2]

[1]U.S. Department of Energy (DOE) Great Lakes Bioenergy Research Center and [2]Department of Bacteriology, University of Wisconsin, Madison; [3]Laboratory of Genetics, University of Wageningen, Wageningen, The Netherlands; [4]Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South Africa; [5]DOE Joint Genome Institute, Walnut Creek, California; and [6]DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico

Exploring natural niches rich in recalcitrant plant biomass may represent a promising avenue for the discovery and characterization of microorganisms with novel ligno-cellulolytic capabilities. Fungus-growing termites (Macrotermitinae) are major decomposers in tropical areas of the Old World (Asia and Africa), where they form some of the most complex colony and mound structures of any insect group. Although other symbiotic relationships have played an essential role in termite evolution (including intestinal microbes), only the Macrotermitinae have evolved a mutualistic association with fungi of the genus *Termitomyces* (Tricholomataceae, Basidiomycotina). The fungus aids in the degradation of plant material, and is housed in a special structure in the nest, the fungus comb, which is maintained by the termites through the continuous addition of predigested plant substrate. The termites themselves play a crucial role in this substrate preparation, because all forage material passes through the termite gut before incorporation

in the fungus comb. Here we explore whether bacteria with ligno-cellulolytic properties are present in guts (potentially aiding predigestion) and in the fungus comb (where the predigested substrate is utilized by the mutualistic fungus). We show using targeted microbial isolations that bacteria with ligno-cellulytic properties can readily be obtained from fungus-growing termite nests. Further, we characterize the abilities of a subset of these microbes to degrade lignin and cellulose, and present the draft genome of a candidate lignin-degrader in the genus *Sphingomonas*. Our findings suggest that the fungus-growing termite symbiosis may provide a promising niche for microbes with novel ligno-cellulolytic properties, and future plans in the system include exploring the bacterial communities associated with comb and gut environments of the South African termite *Macrotermes natalensis* through metagenomic analyses.

Comparative Genomics of the *Pleurotus ostreatus* Genome

Francisco Santoyo,* Lucía Ramírez (lramirez@unavarra.es), and Antonio G. Pisabarro

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

In this poster we describe three studies dealing with the comparison of the genome sequences determined for the two protoclones (haplotypes) of the dikaryotic basidiomycete *Pleurotus ostreatus*.

1. Genome wide screening for helitrons: Helitrons are a new class of transposons that have been found in all eukaryotic kingdoms. Their copy numbers are, however, highly variable, even when closely related species are compared. The genes present in the helitron sequence code for Rep/helicase-like and replication protein A (RPA)-like proteins suggesting that these genetic elements transpose by a rolling-circle mechanism; yet, this hypothesis has not been supported by experimental evidence up to now. The main difficulty for identifying helitrons in genome sequences derives from their scarce and tiny structural features. Helitrons are characterized by a 5′ TC terminus and a 3′ CTRR terminus and their sequence includes a predicted small hairpin structure near the 3′ CTRR end. They are found to be preferentially inserted into the dinucleotide AT. Some elements encode Rep/helicase-like and RPA-like proteins that may be involved in the transposition process. The elements that encode Rep/helicase are considered putative autonomous elements.

In order to screen the genome of *P. ostreatus* for the presence of helitrons, we have used the Helsearch program (http://www.pnas.org/content/106/31/12832.full) trough the genome sequence assemblies of the two *P. ostreatus* monokaryons. With this software 11 and 9 putative helitrons were found in the PC15 and PC9 monokaryons respectively. Only five of them were found to be present in both protoclones. We have also performed a helitron search in the *Tremella mesenterica* genome and we have found 20 helitrons, all of them different to those found in the genome of *P. ostreatus*.

2. Haplotype unique genes: In order to perform a whole transcriptome analysis using the Solid ultrasequencing platform we have determined the unique genes in PC9 and PC15 protoclons to facilitate the allocation of the sequence reads to the two sequenced haplotypes. This study has revealed that 717 gene models were unique to PC9 and 327 to PC15.

3. Comparison of versions 1 and 2 of the PC15 assemblage: Early this year, the PC15 Version 2 genome assembly was finished. To compare the two assembly versions we proceed to a synteny comparison. The genome is now very complete. Seven scaffolds are

complete as chromosomes, telomere to telomere. From the comparison between the two assemblies the main differences are located in scaffolds 2 and 6. The old scaffold 12 is now included in the new scaffold 11 and one 6kpb duplication of the version 1 mitochondrial genome has been eliminated.

Resequencing a Wide-Range of Bioenergy-Relevant Species via Short Read Technology; SNPs, Indels, Structural Variation, and Population Analysis

Wendy Schackwitz* (WSSchackwitz@lbl.gov), Joel Martin, and Len A. Pennacchio

DOE Joint Genome Institute, Walnut Creek, California

We have completed resequencing studies using the Illumina platform on 125 individual genomes covering inbred laboratory strains as well as wild isolates, ranging in complexity from individual haploid prokaryotes and fungi, diploid plants, and evolving populations. The process of analysis for resequencing projects at the JGI will be described in detail.

New DNA Polymerases Improve Fidelity, Specificity and Reliability of PCR

Thomas Schoenfeld, Nicholas Hermersmann, Krishne Gowda, Darby Renneckar, Michael Moser, and **David Mead*** (dmead@lucigen.com)

Lucigen Corporation, Middleton, Wisconsin

DNA amplification is integral to most methods of nucleic acid sequencing, detection and analysis. Amplification errors and bias due to misincorporation, inefficient synthesis through difficult templates (e.g. GC or structure-rich) and/or background amplification significantly compromise the quality of next-gen sequence data. Using high-throughput screens, directed mutagenesis, protein engineering and buffer optimization, we are developing a suite of novel thermostable DNA polymerase that address limitations in the available amplification systems. Voodoo DNA polymerase is highly effective at amplifying otherwise intractable seqeunces and is resistant to impurities, e.g. blood component or salts, simplifying sample preparation. A novel proofreading DNA polymerase, PyroPhage HiFi Pol, allows the highly accurate synthesis needed for deep sequencing projects. Another new DNA polymerase, ColdStart Pol has inherently low activity at room temperature that enables hot start PCR without the cost or compromise of standard antibody or chemical-based hot start methods.

Genome Sequences of the Unicellular, N_2-Fixing Cyanobacterial Genus *Cyanothece*

Louis A. Sherman[3]* (lsherman@purdue.edu), Anindita Banerjee,[1] Jana Stöckel,[1] Lawrence Page,[1] Xueyang Feng,[2] Bing Wu,[2] Yinjie Tang,[2] Hongtao Min,[3] Sujata Mishra,[3] Xiaohui Zhang,[3] Stephen J. Callister,[4] Richard D. Smith,[4] and Himadri B. Pakrasi[1,2]

[1]Departments of Biology and [2]Energy, Environmental and Chemical Engineering, Washington University, St. Louis, Missouri; [3]Department of Biological Sciences, Purdue University, West Lafayette, Indiana; and [4]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington

Cyanothece are unicellular, diazotrophic cyanobacteria with a versatile metabolism and very pronounced diurnal rhythms. Since nitrogen fixation is exquisitely sensitive to oxygen, *Cyanothece* species utilize temporal regulation to accommodate these incompatible processes in a single cell. When grown under 12h light-dark (LD) periods, they perform photosynthesis during the day and N_2 fixation and respiration at night. During this process, carbohydrates and amino acids are compartmentalized in granules in the light and dark, respectively. In essence, *Cyanothece* creates an O_2-limited intracellular environment to perform oxygen-sensitive processes such as N_2-fixation and H_2 production. Some strains also grow exceedingly well on glycerol. The excellent synchrony of a culture under LD diazotrophic conditions permits analysis of cellular morphology, mRNA levels, proteomics and metabolomics as a function of time.

Complete genome sequences of six *Cyanothece* strains are now available (ATCC 51142, sequenced at the Washington Univ. Genome Center; and PCC 7424, PCC 7425, PCC 7822, PCC 8801, and PCC 8802, sequenced by the DOE Joint Genome Institute). The sequences reveal significant metabolic diversity within this group of cyanobacteria. The genome sequence information is being used to generate a *Cyanothece* pan-genome (in collaboration with JGI), comprising the "core genome" (containing all of the genes common to each genus member) and the "dispensable genome" (containing unique genes or genes shared between two or more strains). The unique genes are likely to confer strain-specific attributes and will be analyzed for their role in functions such as H_2 production since the *Cyanothece* evolve large quantities of H_2.

In preparation for conducting relative quantitative proteomics analyses, using the AMT tag proteomics approach, construction of reference peptide databases for 6 of 7 *Cyanothece* strains has been completed. Over 460 LC-MS/MS datasets have been generated and analyzed using the high-throughput proteomics capabilities at PNNL. Complete databases correspond to strains *Cyanothece* sp. PCC 8801, PCC 8802, PCC 7424, PCC 7425, PCC 7822, and ATCC 51142. Percent observed coverage of predicted proteins from unique peptides (10% false discovery rate) ranges from roughly 40% to 70%, which inversely correlates to the size range of genome sequences; *Cyanothece* sp. PCC 7424 has the largest genome sequence (~6.5 Mb) and smallest percent observed coverage among the strains.

We are most interested in the key metabolic processes involved in nitrogen fixation, hydrogen production, photosynthetic and respiratory electron transport, glycerol utilization and CO_2 capture and concentration. We will highlight the major differences among the 6 *Cyanothece* strains and discuss the metabolic properties that will make one or more strains a good model organism for energy production (e.g., hydrogen production, biodiesel formation) and CO_2 capture. This project was funded by a series of grants through DOE including a Grand Challenge in Membrane Biology through EMSL at PNNL and a hydrogen grant from GTL.

Genome Improvement of DOE JGI Flagship Plant Genomes

David Sims* (dsims@hudsonalpha.org), Jane Grimwood, and Jeremy Schmutz

DOE JGI HudsonAlpha Genome Sequencing Center, Huntsville, Alabama

The DOE JGI is placing special emphasis on improving the accuracy and completeness of plant genomes that have been selected because of their DOE mission-relevance and economic importance. These "flagship" plants are switchgrass, sorghum, poplar, soybean, foxtail millet, physcomitrella, and chlamydomonas. Unlocking the genetic information from these plants has the potential for facilitating improved crop yields, disease and insect

resistance and drought tolerance for the crops, as well as better understanding of oil synthesis and cell walls for biofuel feedstocks. By sequencing the DNA of these plants, DOE JGI is making significant contributions toward meeting the world's energy needs.

Typically, the initial draft of the plant genomes is carried out at the JGI in Walnut Creek. Genome improvement, for flagship genomes, takes place at the Genome Sequencing Center of HudsonAlpha Institute for Biotechnology in Huntsville, AL. Assembly and annotation of the sequence is a collaborative effort between the JGI Phytozome Annotation group and HudsonAlpha. The sequence and analysis is made public through the phytozome web site at www.phytozome.net.

Currently, plant whole genomes are drafted using 3 and 6 kb plasmid, fosmid and BAC Sanger libraries. Following the initial assembly of the draft end-reads, the customary protocol for plant genome improvement at HA begins with the selection of targeted areas of the genome. The target area is isolated as a subset of the whole to allow for a more workable subproject size. The subprojects range from either fosmid- or BAC-size to 2 MB. Targets may be selected for a variety of reasons that include gene-rich areas, QTL regions, or a region with significant number of gaps in the sequence and unresolved repetitive sequence.

The objective is to completely (or as near as possible) resolve each base call in the target subproject and then incorporate that "complete" sequence back into the entire genome assembly. Methods for making this improvement include primer walks with a variety of chemistries and templates, transposons and shatters of clones (the latter two techniques being essentially subprojects of the subprojects). Efforts are currently underway to explore practical uses of next-generation sequencing technologies for the extremely large and multi-copied plant genomes.

New public releases of plant genomes occur periodically for each flagship genome, as significant iterative improvements are made to the entire sequence. The DOE JGI strives to provide a greater understanding of the genetic variation and functional adaptations of flagship plant genomes, thus improving our nation's biofuel capabilities.

Comparative Genomics of Various Lineages of the Dry Rot Fungus *Serpula lacrymans*

Inger Skrede[1]* (inger.skrede@bio.uio.no), Mikael Brandström Durling,[2] Jan Stenlid,[2] Håvard Kauserud,[1] and Nils Högberg[2]

[1]Department of Biology, Microbial Evolution Research Group, University of Oslo, Oslo, Norway and [2]Department of Forest Mycology and Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden

The dry rot fungus *Serpula lacrymans* is the most damaging destroyer of wooden materials in houses in the temperate regions. *Serpula lacrymans* var. *lacrymans* is attacking houses all over the northern hemisphere, but is only found in a few locations in nature. In contrast, the closely related *S. lacrymans* var. *shastensis* has only been found in nature and has apparently a limited distribution in western North America. Two monokaryons of var. *lacrymans* from the same spore family have recently been genome sequenced by DOE Joint Genome Institute using Sanger sequencing and JGI is now in the process of sequencing var. *shastensis* using 454 pyrosequencing. In addition, we plan to sequence the genomes of 20 dikaryotic *Serpula* strains from various lineages using the Illumina sequencing technique and assemble these data using the already sequenced genomes as a guide. These genomic data will be used to study various topics, including (1) Which

genomic changes have happened during the transition from growth in nature to buildings in *Serpula lacrymans*? (2) Are there distinct genomic differences between different invasive populations of var. *lacrymans*? (3) Characterize mating type and vegetative incompatibility loci, and investigate how the evolution of MAT and *vic* loci will be affected by bottlenecks and invasiveness. During the project we will also gain experience in how to work with dikaryotic mycelia in this context.

Single Cell Genomics Center at Bigelow Laboratory for Ocean Sciences

R. Stepanauskas* (rstepanauskas@bigelow.org), M.E. Sieracki, D. Masland, N. Poulton, M. Martinez-Garcia, B. Swan, B. Tupper, W. Korjeff-Bellows, and M. Lluesma

Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine

The Bigelow Single Cell Genomics Center (SCGC) was established in 2009 as the first of its kind shared-user facility (www.bigelow.org/scgc). The SCGC aims at making the single cell genomics technology available to the broad scientific community, serving as an engine of discoveries in the areas of microbial ecology, evolution, and bioprospecting. The SCGC offers several services on per-fee basis, including cell isolation, whole genome amplification, PCR-based screening, and sequencing of PCR products.

For single cell isolation, the SCGC utilizes high-speed, jet-in-air research flow sorters. The whole genome amplification (WGA) and polymerase chain reactions (PCR) are performed in 384-well microplate format employing robotic liquid handling. All WGA and PCR reactions are monitored real-time for cherry picking and as an early QC measure. Elaborate efforts are taken to prevent DNA contamination during cell sorting and WGA setup. DNA sequencing is not performed at SCGC, but is rather accomplished through collaborations or outsourcing to established sequencing centers. A sophisticated laboratory information management system is being implemented to track and store project and sample meta- and analytical data. Currently, SCGC has the capacity to perform sorting and WGA on about 10,000 individual cells per week.

The SCGC is currently utilized by multiple in-house and external research projects, spanning genomic studies of the uncultured prokaryotes, protists and viruses from marine, freshwater, and subsurface environments. Major institutional collaborations include JGI (the GEBA project) and Genoscope (Tara Oceans). Research result highlights include findings of inorganic carbon fixation genes in mesopelagic bacterioplankton, identification of key carriers of rhodopsin and bacteriochlophyll genes in freshwater lakes, and the recovery of novel Archaea and Bacteria phyla from terrestrial subsurface.

Deep Sequencing of Ribosomal RNA Genes During an Algal Bloom in a Eutrophic Lake: A Primer for Metagenomic Sequencing

Blaire Steven* (bsteven@uwyo.edu) and Naomi Ward

Molecular Biology, University of Wyoming, Laramie

Freshwater algae (both true algae and cyanobacteria) are a potentially rich source of biomass feedstock for biodiesel and hydrogen, as well as animal feed supplements, soil amendment materials, and other commercial products. They also serve as an environmental carbon dioxide sink that could be exploited through geoengineering.

Algaculture in open ponds is less expensive than closed-reactor growth, but is vulnerable to the activities of other pond microorganisms. There is an incomplete understanding of the contribution of non-algal microorganisms to algal population dynamics. The locally concentrated algal populations seen in algaculture and algal geoengineering are also observed during naturally occurring algal bloom events. The latter are usually perceived as an entirely negative consequence of eutrophication, resulting from increased nutrient input into a water body. However, they provide ideal conditions for studying bacterial-algal interactions relevant to algaculture and algal geoengineering. We are pursuing high-throughput sequencing approaches to characterize microbial communities associated with freshwater algal blooms. This work is innovative because there are no previous reports of metatranscriptomic approaches to understanding bacterial-algal interactions in freshwater systems. The expected outcome of the work is a resource that the scientific community and industrial partners can use to better understand the organisms and processes involved in development and decline of freshwater algal populations. This improved understanding will be applicable to the management of both beneficial (e.g. algal biofuel production) and harmful (e.g. harmful algal bloom) aspects of algal growth.

In order to optimally select sampling time points for metagenome/metatranscriptome sequencing in 2010, we conducted a comprehensive rRNA-based survey of microbial community composition changes over the course of a 2009 algal bloom in Labonte Lake, Wyoming. Samples from three lake habitats were collected: sediment, water column, and macroscopically visible algal mat material. Samples were subdivided for water quality analysis, microscopy, and DNA extraction. Analysis of dissolved nitrogen and phosphorus levels indicated that the lake was highly eutrophic. Using high-throughput pyrosequencing, we generated 141,155 bacterial 16S rRNA gene sequences, and archaeal 16S rRNA gene sequencing is underway. Preliminary analysis suggested that there were both spatial differences and temporal changes in microbial diversity associated with the bloom event. For example, *Actinobacteria* were dominant in water column samples before the bloom but were replaced by *Betaproteobacteria* during the peak bloom and into the bloom decline. The algal mat bacterial community was dominated by four phyla: *Cyanobacteria*, *Alphaproteobacteria, Betaproteobacteria,* and *Bacteroidetes*. As anticipated, 16S rRNA gene sequences related to the *Cyanobacteria* were numerically dominant. Among the non-photosynthetic bacteria in the algal mat, *Alphaproteobacteria* were most abundant.

In summary, our initial results indicated that specific lineages of bacteria might be associated with different developmental stages of the algal bloom. In the summer of 2010 we will collect samples from algal blooms in Labonte Lake for metagenomic and metatranscriptomic sequencing by the JGI Community Sequencing Program, through support for a project entitled “Metatranscriptomic analysis of bacterial-algal interactions: an ecological foundation for enhancing algal biofuel and geoengineering initiatives”. The result of this study will allow us to identify key organisms and processes that contribute to algal bloom dynamics.

Reclassification of *Cellvibrio gilvus* ATCC 13127 to the Genus *Cellulomonas* using Whole-Genome Sequence Analysis

Garret Suen[1]* (gsuen@wisc.edu), Frank O. Aylward,[1] Joseph A. Moeller,[1] A. Christine Munk,[2] Lynne Goodwin,[2] Cameron R. Currie,[1] David Mead,[1,3] and Phil Brumm[1,3]

[1]Great Lakes Bioenergy Research Center, University of Wisconsin, Madison; [2]DOE Joint Genome Institute, Walnut Creek, California; and [3]Lucigen and C5-6 Technologies, Middleton, Wisconsin

New and improved biomass-degrading enzymes are a major requirement to achieving a viable cellulosic biofuels business. As part of the Department of Energy's Great Lakes Bioenergy Research Center, C5-6 and Lucigen have created a functional screening program for discovering and developing biomass-degrading enzymes from both known and novel cellulolytic organisms. One such organism is *Cellvibrio gilvus* ATCC 13127, originally isolated from the feces of bovines. To better understand this organism, a draft genome for *Cellvibrio gilvus* was sequenced at the DOE's Joint Genome Institute from constructed clone libraries. Analysis of this draft revealed surprising results, which placed into question this organism's taxonomic identity. The initial placement of this organism within the genus *Cellvibrio* was originally determined based on morphological comparison using *Bergey's Manual* in the 1950s. Since then, no further analyses have been performed to confirm its taxanomic placement. Using the draft genome sequence of this organism, we have discovered that it does not belong to the Gammaproteobacteria, but rather, belongs to the genus *Cellulomonas* in the phylum Actinobacteria.

We present multiple lines of evidence in support of this proposed taxonomic change, including 16S rDNA phylogenetic analysis, multi-locus sequencing, whole-genome taxonomic distribution analysis, *in silico* DNA-DNA hybridization data, GC content analysis, and Gram-staining. Our 16S rDNA and multi-locus sequence analysis definitively places this organism within the phylum Actinobacteria, specifically in the genus *Cellulomonas*. Whole-genome taxonomic distribution analysis of this organism's predicted proteome shows that 74% of its proteins have their closest matches to proteins belonging to Actinobacteria. Further *in silico* DNA-DNA hybridization analysis found that it has a significantly closer hybridization value to the genomes of *Cellulomonas fimi* and *Cellulomonas flavigena* than to *Cellvibrio japonicus*. GC content analysis of these genomes revealed that *C. gilvus* has a GC content of 73.7%, similar to the ~75% GC content for both *Cellulomonas fimi* and *Cellulomonas flavigena*. In contrast, the GC content for *Cellvibrio japonicus* is 52%. Finally, Gram-staining revealed a Gram-positive identity, as expected for organisms in the Actinobacteria. Based on these data, we propose the reclassification of this organism as *Cellulomonas gilvus* nov. comb. (type strain ATCC 13127^T).

Having confirmed the taxonomic identity of *Cellulomonas gilvus*, we also present preliminary data describing the CAZy gene repertoire of this cellulolytic organism.

Genomic Analysis of *Fibrobacter succinogenes* Reveals an Enigmatic Cellulose Degrader

Garret Suen[1]* (gsuen@wisc.edu), David M. Stevenson,[2] Hazuki Tashima,[3] Lynne A. Goodwin,[3] Cameron R. Currie,[1] Paul J. Weimer,[2] David Mead,[4] and Phil Brumm[4]

[1]Great Lakes Bioenergy Research Center, University of Wisconsin, Madison; [2]USDA-Agricultural Research Service, Dairy Forage Research, Madison, Wisconsin; [3]DOE Joint Genome Institute, Walnut Creek, California; and [4]Lucigen and C5-6 Technologies, Middleton, Wisconsin

Fibrobacter succinogenes is a highly-cellulolytic bacteria found in the rumen of bovines. It is a gram-negative bacterium that belongs to the phylum Fibrobacteres and is an enigma with regards to how it adheres to and degrades cellulose. It is one of the most abundant and active cellulolytic bacteria in the rumen ecosystem, but enzymes isolated from *F. succinogenes* that are capable of hydrolyzing native crystalline cellulose at a rapid rate have not been identified. In order to better understand plant cell wall degradation by this organism, we have developed new tools to capture, express, and identify many of the carbohydrate-active enzymes (CAZymes) from this microbe. This includes a genome sequence for *F. succinogenes*, which was recently completed by the DOE Joint Genome Institute.

Our analysis of the *F. succinogenes* genome confirms its phylogenetic placement as the only sequenced representative of its phylum, the Fibrobacteres. A taxonomic distribution analysis of its predicted proteome reveals that it has a large number of proteins closely related to bacteria in the phylum Bacteroidetes and Firmicutes. Based on metabolic pathway reconstruction of *F. succinogenes*, this organism appears to lack catabolic pathways for fatty acids and many amino acids, confirming the observation that carbohydrates are its sole energy source. Further analysis reveals enzymes present for glycolysis and the presence of an incomplete TCA cycle, thereby accounting for the fermentation products produced by the organism. This organism also lacks many pentose phosphate pathway enzymes, explaining why *F. succinogenes* can degrade, but not metabolize xylan. This supports previous findings that *F. succinogenes* grows almost exclusively on glucose, cellobiose and cellodextrins, and likely uses its xylan-degrading machinery to gain access to cellulose in plant cell walls. This organism also has a complete set of enzymes for the biosysnthesis of valine, leucine and isolecucine, suggesting its growth requirement for branched-chain volatile fatty acids is not linked to amino acid biosynthesis but possibly to fatty acid biosynthesis.

Preliminary analysis of the predicted CAZymes in its genome indicates that it contains 102 glycosyl hydrolases, 12 polysaccharide lyases, 15 carbohydrate esterases and 63 carbohydrate binding modules, one of the highest of any sequenced microbe when expressed as a percent of the total number of proteins. Only 35 of these CAZYmes are annotated for cellulose degradation, and the majority of the CBMs are specific to xylan and other carbohydrates, with only a handful of cellulose binding CBMs. Furthermore, it lacks dockerin and cohesin sequences, does not encode any known processive cellulases, and most of its endoglucanases do not contain carbohydrate binding modules. Therefore, it appears that *F. succinogenes* has a novel mechanism of cellulose degradation. One goal of our CAZyme work is to functionally characterize all predicted glycosyl hydrolases from *F. succinogenes* to determine how well bioinformatic analysis serves as a proxy for determining enzymatic activity. Additional details are provided in the companion poster *Functional Annotation of* Fibrobacter succinogenes *Carbohydrate-active Enzymes*. Another goal is to identify and characterize the unknown proteins involved in binding and degradation of crystalline cellulose.

Development of High Throughput Processes for Constructing Illumina Libraries

Eric Tang[1]* (ETang@lbl.gov), Chris Hack,[1] Shweta Deshpande,[1] Susan Lucas,[2] Jan-Fang Cheng,[1] and the JGI Production Sequencing Group

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

As the demand of constructing Illumina libraries increases, we have started to modify the library construction protocol to adapt the use of multichannel pipette and 96-well plates. With the few simple modification steps, we have doubled the library production efficiency. These modifications include the shearing of DNA with Covaris E210, and the cleaning of enzymatic reactions and fragment size selection with SPRI beads and a magnetic plate holder. We have also designed a set of molecular barcodes to enable the sequencing of many libraries in parallel. The requirements of these barcodes include 4 bases, balanced GC, and at least 2 bases difference between barcodes. The barcode is attached to the adaptor so it does not require third sequencing primer and the barcoded library can be run on the same flowcell/run with other non-barcoded libraries. We have begun to assess the ability to assign reads and the potential bias towards certain barcodes after pooling different number of libraries.

We have recently programmed the Biomek FX robot to carry out the library construction process. Although this process still require manual transfer of plates from robot to other work stations, the processing of 96 Illumina libraries takes approximately 6 - 8 hours. This semi-automated process represents a significant increase of library capacity comparing to the manual process. We will present the progress and the challenges of these scale-up processes.

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

JGI Innovative Seed Program Brainstorm

Kristen Taylor* (kmtaylor@lbl.gov), Henrik Nordberg, David Gilbert, and Tatyana Smirnova

DOE Joint Genome Institute, Walnut Creek, California

Do you have a great idea for the JGI? Is there something you wish the JGI would do using a new technology? The Innovative Seed Program (ISP) was put in place to connect good ideas to the right people. The ISP hosts monthly meetings where ideas are presented and collected in a collaborative open forum. Ideas are entered in a tracking system and an advocate is assigned to help find the right area at the JGI where that idea could be fostered. As users of the JGI websites and tools, you are an excellent source of new ideas and the ISP wants to hear from you. Advocates will be available at this poster station to collect your ideas and to show you where to submit future ideas.

Natural and Induced Variation in *Brachypodium distachyon* Cell Walls

Ludmila Tyler[1,2]* (ltyler@berkeley.edu), Michael A. Steinwand,[2] and John P. Vogel[2]

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

Biomass from grasses grown as dedicated energy crops (*e.g.* switchgrass and *Miscanthus*) and residues from grain crops (*e.g.* corn, wheat, and rice) are poised to become significant feedstocks for the emerging cellulosic biofuel industry. Current production platforms convert sugars locked in the cellulose and hemicellulose portions of the cell walls into biofuels (ethanol, butanol, *etc.*) through fermentation. Yet, relatively little is known about the genes that control the composition and structure of the unique grass cell wall and the ways in which characteristics of the cell wall can be modified to improve biofuel production. The fastest way to gain this information is through the use of an appropriate model system. The recent annotation of the *Brachypodium* genome supports its use as a general model for the grass cell wall, because the complement of genes putatively involved in cell-wall metabolism was markedly similar in *Brachypodium*, rice and sorghum. To screen for compositional and/or structural variation in cell walls, we are examining *Brachypodium* stems with near-infrared spectroscopy (NIRS). This method has revealed extensive natural variation within a collection of over 170 genetically diverse inbred lines. Additionally, NIRS has been used to identify cell wall mutants from ethyl-methane-sulfonate- and fast-neutron-mutagenized populations of *Brachypodium*. An overview of the annotation of glycoside hydrolase genes implicated in cell-wall-metabolism, a summary of the screening for both induced and natural variation in cell-wall composition, and initial results from fermentation tests will be presented.

Completion of the *Brachypodium distachyon* Genome Project

John Vogel[1]* (brachypodium@gmail.com), David Garvin,[2] Todd Mockler,[3] Jeremy Schmutz,[4] Daniel Rokhsar,[5] Michael Bevan,[6] and the International Brachypodium Initiative[7]

[1]USDA-ARS Western Regional Research Center, Albany, California; [2]USDA-ARS Plant Science Research Unit, University of Minnesota, St. Paul; [3]Oregon State University, Corvallis; [4]Hudson Alpha Institute of Biotechnology, [5]DOE Joint Genome Institute, Walnut Creek, California; [7]John Innes Centre, Norwich, United Kingdom; and [7]The full list of participants can be found in Nature 463: 763-768. 2010

To foster the development of *Brachypodium distachyon* (Brachypodium) as a model system to study questions unique to the grasses, a large collaboration of researchers collectively known as the International Brachypodium Initiative published a draft Brachypodium genome sequence and annotation in the February 11, 2010 issue of Nature. The sequence and annotation was of extremely high quality. A key finding was the identification of nested chromosome insertions into centromeric regions as a common mechanism of grass chromosome evolution. Manual examination of 2,755 Brachypodium genes revealed that, overall, Brachypodium gene family structure was very similar to rice and sorghum indicating that Brachypodium can serve as a model for most aspects of grass biology. When combined with the natural attributes and existing genomic resources (e.g. rapid lifecycle, small stature, high-efficiency transformation, simple growth requirements, large germplasm collection, T-DNA mutants, microarrays, BAC libraries etc.) the genome sequence will enable researchers to utilize Brachypodium for a wide array of experimental approaches. An overview of the *Brachypodium* genome project will be presented.

A Comparative Genomics Approach to the Evolution of the Ectomycorrhizal Symbiosis

Benjamin E. Wolfe* (bewolfe@fas.harvard.edu) and Anne Pringle

Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts

Ectomycorrhizal fungi form ectosymbioses on the surface of woody plant roots and play key roles in the functioning of forest ecosystems throughout the world. This symbiotic growth is in contrast to free-living, saprotrophic fungi that obtain carbon from the decomposition of organic matter. Recent whole genome sequence data for ectomycorrhizal fungi has greatly advanced our knowledge of the molecular genetics of the ectomycorrhizal symbiosis. However, we still lack a detailed understanding of the genomic differences between ectomycorrhizal and saprotrophic fungi. We are using the fungal genus *Amanita* for comparative genomics of the ectomycorrhizal symbiosis. Most species in this genus are ectomycorrhizal, but a few species are found in habitats without woody plant hosts and are saprotrophic. We have demonstrated that the ectomycorrhizal symbiosis has evolved once within the clade of mushroom forming fungi that contains the genus *Amanita.* Based on PCR screens*,* we have found genes for several different cellulases in the genomes of saprotrophic *Amanita* species, but we have not found these genes in ectomycorrhizal *Amanita* genomes. The whole genome of a free-living species, *Amanita thiersii,* is currently being sequenced at the Joint Genome Institute. This genome sequence, in combination with sequencing of ectomycorrhizal *Amanita* genomes, will allow us to answer the following questions: 1) What is the structure and function of the enzyme system that *Amanita thiersii* uses to decompose cellulose in grass litter? 2) Are genes involved in decomposition clustered within the genome of *Amanita thiersii*? 3) What are core genes and genomic features that are shared by all *Amanita* species and what genes and genomic features distinguish ectomycorrhizal species from saprotrophic species?

Proteomics Analysis of Lipid Bodies and Associated Endomembranes of the Haptophyte Algae *Emiliania huxleyi* and *Isochrysis galbana*: Clues to the Biosynthesis and Function of Polyunsaturated Long-Chain Alkenones, and Potential for Biofuels

Gordon Wolfe* (GWolfe2@csuchico.edu) and William Erlendson

Department of Biological Sciences, California State University, Chico

Emiliania huxleyi and some related hapotphyte algae produce as neutral lipids a set of Polyunsaturated long-chain (C_{37-39}) Alkenones, Alkenoates, and Alkenes (PULCA). These biomarkers are widely used for paleothermometry, but the cellular machinery of these unique lipids remains largely unknown. Eltgroth et al. (2005) previously showed that these taxa, like other eukaryotes, package their neutral lipid into cytoplasmic lipid bodies (LBs). We developed a method for fractionation and separation of PULCA-rich fractions from *Emiliania* and its sister taxon *Isochrysis*, and present here their LB proteomes. In both taxa, proteins associated with LBs include functional classes found in LBs of other organisms, including acyl and prenyl lipid biosynthetic and catabolic enzymes, as well as associated structural and transport proteins. Unlike LBs in plants and animals, haptophyte LBs do not appear to have unique packaging proteins, but we found numerous PAP-fibrillin homologs. Other LB-associated proteins from plastids, peroxisomes and mitochondria, and endomembranes likely hint at the cycling of these quasi-organelles through anabolic and catabolic phases. LB and endomembrane acyl and prenyl enzymes

are both implicated in PULCA biosynthesis, and our findings partially support the biosynthetic pathway hypothesized by Rontani et al. (2006) based on structural analysis of this lipid family. However, our results support the involvement of both plastidial fatty acid biosynthesis and cytoplasmic fatty acid type-I or polyketide synthase systems, with polydesaturation possibly provided by prenyl-lipid-derived machinery, and we suggest that regulation of acyl-CoA and ACP pools may help regulate synthesis of different lipid pools in these organisms. We also found massive stimulation of PULCA and LBs when cells were given excess bicarbonate under N or P limitation, which may have biofuel potential, but that these taxa use the LBs in somewhat different ways, which likely reflects the complex life cycle of haptophytes. In all life stages, these lipids serve as energy stores, well supported by the finding of both anabolic and catabolic pathways in our LB proteome. In non-motile diploid *Emiliania* cells, PULCA may also play a role in calcification: we observed PULCA associated with the cell wall, machinery associated with secretion in our LB proteome, and obtained co-stimulation of calcification, PULCA and LB in some strains. In contrast, in motile haploid *Isochrysis,* LBs show a much more complex proteome and may possibly as a form of eyespot apparatus for phototaxis. To our knowledge, this is the first proteomics study of chromists and of this globally important alga.

De Novo Sequencing Strategy for Daunting Genomes

Cheng-Cang Wu, Rosa Ye, Svetlana Jasinovica, Megan Wagner, Ronald Godiska, and **David Mead*** (dmead@lucigen.com)

Lucigen Corporation, Middleton, Wisconsin

Next generation sequencing technologies can rapidly and economically produce a draft genome of an organism de novo. However, the quality of the draft data is seldom more than 80% complete with >10e5 contigs for large genomes, which is insufficient for many applications. More importantly no hands-on genomic resources (e.g. sequenced BACs, BAC physical maps) are produced for functional genomics after sequencing. Sequence data that is closer to 95% finished with the unambiguous order and placement of genes would have the greatest utility for scientific and commercial research. Tools that bridge the gaps between massively parallel short read sequencing technologies (35-500 bases) and the need for large scaffolds to accurately assemble complex repeat rich genomes (>100,000 bases) are needed. Lucigen has successfully developed new tools and methods for the construction of large insert random shear BAC libraries. This advance allows the production of whole genome libraries that are unbiased by the non-random distribution of restriction enzyme sites, which significantly reduces the number of clones needed to finish a genome while eliminating gaps due to sequence bias. To date Lucigen has constructed more than 100 random shear BAC libraries of microbes, plants and animal species for researchers around the world. Combining random shear BAC library capabilities with next generation sequencing technologies should theoretically result in nearly complete coverage, assembly of even daunting genomes and useful genomic information and resources simultaneously. If successful, this approach could allow the rational sequencing and analysis of daunting genomes such as wheat and loblolly pine, enable complete coverage of complex metagenomes and simplify the resequencing of repeat rich regions of the human genome such as the major histocompatibility complex. We will present a strategy for achieving this aim that seeks to provide a nearly finished genome while reducing computational complexity, maximizing the efficiency of the existing technologies and produce more useful de novo reference genomes.

Identify Novel Phylogenetic Markers for the Archaea and Bacteria Genomes

Dongying Wu[1,2]* (DYwu@lbl.gov) and Jonathan A. Eisen[1,2]

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

For phylogenomic and metagenomic studies of bacteria and archaea, more phylogenetic markers are needed in addition to the small subunit rRNA gene and the handful of proteins markers current in use (such as recA and rpoB).

To identify more phylogenetic markers for bacteria, we've built gene families for 85 bacterial representative genomes. The 85 organisms were selected so that phylogentic diversities (PD) are maximized. The PD calculation was based on a maximum likelihood tree from the concatenated alignments of 31 markers in AMPHORA. We've identified 25 new phylogenetic markers candidates as a result. The 25 new marker candidates are selected because they are evenly distributed across the genomes and each genome only has a single copy of the gene. Phylogenetic tree building and tree topology comparisons indicate that the 25 novel markers are as good as the 31 AMPHORA markers to study the bacterial phylogenies.

We've established a protocol to identify automatically phylogenetic marker candidates for any given phylogentic groups. The protocol uses BLAST and MCL clustering algorithms to generate gene families for a given group of genomes. Phylogenetic trees are built for the gene families and clades from the trees are automatically sampled and evaluated for universality and evenness in terms of their distributions. HMM profiles are built for the clades with genes distributed across the organisms with a single copy in each genome. HMM searching against the entire proteome of the group is applied to evaluate how distinct the gene families are. We've build distinct single-copied gene families that evenly distributed within a phylogenetic group (we've studied the archaeal domain and 10 bacterial phyla). HMM profiles were built for 5189 families that can be potential markers for the lineages of interest. Clustering and tree building analysis of the consensus sequences from all the families reveals that we can identify 62 gene markers that each span archaea and at least 4 bacteria phyla, as well as 324 bacterial gene markers that each covers at least 5 phyla. We are in the process of studying the distribution of the ~300 marker candidate across all sequenced genomes as well as the topologies of the phylogenetic tree built from them to establish a marker database for phylogenomic and metagenomic studies.

Single Cell Genome Size Estimation: The Single Copy Gene Approach

Chi Yang,[1,2] Cliff Han,[1] Ramunas Stepanauskas,[3] Tanja Woyke,[4] Chuan-Hsiung Chang,[2] and **Gary Xie**[1]* (xie@lanl.gov)

[1]DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico; [2]National Yang-Ming University, Taiwan; [3]Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine; and [4]DOE Joint Genome Institute Production Genomics Facility, Walnut Creek, California

Fixation-Free Fluorescence *in situ* Hybridization for Targeted Enrichment of Microbial Populations

Suzan Yilmaz[1]* (syilmaz@lbl.gov), Mohamed F. Haroon,[2] Brian A. Rabkin,[1,3] Gene W. Tyson,[2] and Philip Hugenholtz[1]

[1]Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, California and [2]Advanced Water Management Centre, The University of Queensland, Australia; [3]present address: Illumina Corp., Hayward, California

Fluorescence *in situ* hybridization (FISH) is a powerful method for visualizing microbial populations in a community setting and can be used as the basis for separating targeted populations from a multi-species background through fluorescence activated cell sorting (FACS). Cell fixation is an accepted and reportedly integral part of the FISH protocol that serves to stabilize cell integrity and facilitate probe permeability for imaging. For microorganisms, this is most commonly achieved with paraformaldehyde (PFA) which crosslinks proteins and nucleic acids. However, for genomic or proteomic applications, unfixed biomass is desirable to minimize possible sequencing biases introduced by crosslinking and also to avoid possible cell lysis difficulties. In this study, we made a simple modification to the standard ribosomal RNA-targeted FISH protocol by removing the PFA-fixation and ethanol dehydration steps to allow recovery of unmodified DNA and proteins, and applied broad-specificity FISH probes to a range of environmental samples including termite hindgut and bioreactor communities. Surprisingly, this resulted in no appreciable effect on the quality of FISH images using epifluorescence microscopy. Combined with fluorescence activated cell sorting, Fixation-free FISH facilitates subsequent omic analyses of targeted populations.

Switchgrass Functional Genomics Tools Development

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

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

Switchgrass (*Panicum virgatum* L.) is a perennial C_4 grass species native to North America, which is used as forage on the Great Plains of the USA and has the potential to be a major source of biomass for bio-fuel production. To realize this potential, breeding efforts are underway to improve wild germplasm and forage type switchgrass for bioenergy uses. Development of genomics resources and tools will aid the breeding enterprise, via molecular marker development, and enable forays into molecular and systems biology of this species.

Summer VS16, an upland genotype, and Alamo AP13, a lowland genotype, are the parents of a mapping population that is currently being characterized in the field for traits related to biomass/biofuel production, adaptation, and sustainability. We have clonally-propagated these two genotypes in vitro to generate sufficient genetically-identical plant material for RNA isolation used in various EST projects. EST sequencing of these two genotypes will facilitate SNP discovery for genetic mapping of target traits by switchgrass breeders.

So far, most of EST sequencing effort is concentrated on Alamo AP13, of which the genome is being sequenced, and 7.2 million ESTs using 454 platform have been generated

from multiple cDNA libraries of this genotype. In addition, two cDNA libraries of Alamo AP13 generated from RNA of various organs and developmental stages were constructed and 71,423 end reads were obtained using Sanger sequencing technology. Furthermore, a normalized full length cDNA library, using RNA from a broader range of tissues sources and abiotic stress treatments is in progress and the full length reads will be recovered by combination of 454 and Sanger reads. This will enable further clustering of AP13 cDNA sequences and provide verified full-length clones for functional analysis of proteins in the future.

In parallel, we have generated 1.5 million ESTs from 454 reads of seven Summer VS16 cDNA libraries of roots and shoots harvested at different stages of development to facilitate SNP discovery for the mapping population. The average read length of these ESTs is 202 bp. This dataset has been used as test set for assembly and different assembly programs including NEWBLER, MIRA, TGICL, and PAVE programs have been evaluated. Among these programs, NEWBLER and MIRA gave preferable results in regarding to speed and final assembled data although there still significant problem.

After completion of EST/cDNA sequencing of both genotypes, sequence alignment programs will be used to produce a switchgrass gene index with quality dataset. A switchgrass Affymetrix Genechip will then be designed, followed by development of a reference Gene Expression Atlas. Furthermore, putative SNPs for genetic mapping in populations derived from the two parents will be identified. All these resources and tools will be made available to the broad switchgrass research and breeding communities upon completion.

General Finishing Process with Second-Generation Sequencing Technologies

Lucy (Xiaojing) Zhang* (xiaojing.lucy@gmail.com), Karen W. Davenport, Hajnalka E. Daligault, and **Cliff Han**

Los Alamos National Laboratory, Los Alamos, New Mexico

Our general finishing process has been updated to complement the new sequencing technologies. Sequencing data from 454 standard, 454 paired-end and Solexa libraries are screened according to their quality and randomly picked to provide certain depth of coverage for each different library. The selected data sets are assembled by related assembly software (Newbler for 454 data, VELVET for Solexa data) then overlapping fake reads are generated. New software, which were designed to resolve duplication regions (dup454Finisher and gapResolution) are used to generate specific fake reads after gap closure in the subprojects. The fake reads generated from the Newbler and VELVET assembies and from the two duplication resolving software are assembled together by PhredPhrep. For the remaining gaps PCR and bubble PCR reactions are ordered before passing the project to manual finishing. All these steps are performed as an automated project StartUp process. Manual finishing involves resolving misassemblies, closing the remaining gaps, and increasing the quality of the sequence. The projects go through several manual finishing cycles. The number of the cycles depends on the difficulty of the project (quality of the draft sequence, number of duplications, hard GC stops, GC percentage etc.). The last step in manual finishing is to increase the quality of the sequence. The ‘Polisher’ software was developed to detect discrepancies between the 454 and Solexa data, identify areas of low coverage by Solexa and to pick primers for bubble PCR reactions in low quality regions. Continuous research and development in the wet lab

and in bioinformatics are increasing the quality of finished genomes and reducing the time and money required for finishing.

High EST Coverage Revealed Large Fraction of Alternatively Spliced Transcripts in Fungal Genomes

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

DOE Joint Genome Institue, Walnut Creek, California

Gene modeling has always been a challenge for computational biologists, but it becomes trivial when informed by expressed sequence tags (ESTs). New sequencing technologies such as 454 and Solexa can generate huge number of ESTs, but current programs such as PASA and Newbler are neither fast nor good enough to derive quality gene models from EST sequences. We developed a new algorithm COMBEST that uses EST and genomic sequences as input to generate partial or complete gene models. When applied this algorithm to three genomes—*Chamydomonas reinhardtii*, *Agaricus bisporus*, and *Aspergillus carbonarius*—with varying degree of coverage of 1.7, 22.5, and 51.9 ESTs per kb genomic sequence, we found different fractions of genes with alternative spliced forms of 6%, 15%, and 28% for three genomes respectively. Although the fraction of alternatively spliced genes is an inherent feature of a particular genome and the living condition of the organism harboring the genome, deep EST coverage is clearly essential for revealing the alternatively spliced forms. Since our algorithm also calculates the relative expression level for each splicing isoform, the results from COMBEST can be a useful resource for studying intron slicing and evolution in addition to being a tool for gene modeling in the high-through-put sequencing era. One of the interesting results from our analysis is that the splicing machinery makes all sorts of mistakes at low frequencies.

Attendees

Current as of March 3, 2010

Andrea Aerts
DOE Joint Genome Institute
alaerts@lbl.gov

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

Ed Allen
DOE Joint Genome Institute
eaallen@lbl.gov

Martin Allgaier
Joint BioEnergy Institute
mallgaier@lbl.gov

Eric Alsop
UC Merced
ealsop@ucmerced.edu

Gordon Anderson
Pacific Northwest National Laboratory
gordon@pnl.gov

Iain Anderson
Joint Genome Institute
ijanderson@lbl.gov

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

Frank Aylward
UW-Madison
faylward@wisc.edu

Susan Baldwin
University of British Columbia
sbaldwin@interchange.ubc.ca

Matthieu Barret
BIOMERIT UCC
m.barret@ucc.ie

Adam Barry
DOE Joint Genome Institute
abarry@lbl.gov

Kerrie Barry
DOE Joint Genome Institute
kwbarry@lbl.gov

Khela Baskett
DOE Joint Genome Institute
kbweiler@lbl.gov

John Battista
Louisiana State University
jbattis@lsu.edu

Bonnie Baxter
Westminster College
bbaxter@westminstercollege.edu

Laura Beer
Colorado School of Mines
laurabeer@gmail.com

William Benton
DOE Great Lakes Bioenergy Res. Center
dbenton@glbrc.wisc.edu

Randy Berka
Novozymes, Inc.
ramb@novozymes.com

Stephanie Bernard
Lawrence Berkeley National Lab
smbernard@lbl.gov

Devaki Bhaya
Carnegie Institution
dbhaya@stanford.edu

Christophe Billette
INRA
billette@bordeau.inra.fr

Matthew Blow
DOE Joint Genome Institute
mjblow@lbl.gov

Harvey Bolton, Jr.
Pacific Northwest National Lab
harvey.bolton@pnl.gov

Ben Bowen
Lawrence Berkeley National Lab
bpbowen@lbl.gov

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

Susan Brawley
University of Maine
brawley@maine.edu

Thomas Brettin
Oak Ridge National Laboratory
brettints@ornl.gov

Jim Bristow
DOE Joint Genome Institute
jbristow@lbl.gov

Robert Britton
Michigan State University
rbritton@msu.edu

Igor Brown
SARD/Johnson Space Center
igor.i.brown@nasa.gov

Donald Bryant
The Pennsylvania State University
dab14@psu.edu

Kathy Byrne-Bailey
University of California, Berkeley
k.g.byrne-bailey@berkeley.edu

Fei Cai
DOE Joint Genome Institute
fcai@lbl.gov

Tor Carlsen
University of Oslo
tor.carlsen@bio.uio.no

Patrick Chain
Los Alamos National Laboratory
pchain@lanl.gov

Romy Chakraborty
Lawrence Berkeley National Lab
rchakraborty@lbl.gov

Mia Champion
Broad Institute
championmia@yahoo.com

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

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

Hsin-I Chiang
UC San Diego
hchiang@ucsd.edu

Mansi Chovatia
DOE Joint Genome Institute
mrchovatia@lbl.gov

Marcus Claesson
University College Cork
mclaesson@bioinfo.ucc.ie

Louis Clark
Codexis
louis.clark@codexis.com

Melinda Clark
UC Berkeley
Energy Biosciences Institute
melinda_clark@berkeley.edu

Collin Closek
University of California, Merced
cclosek@ucmerced.edu

Alicia Clum
DOE Joint Genome Institute
aclum@lbl.gov

Attendees

Frank Collart
Argonne National Lab
fcollart@anl.gov

Jackie Collier
Stony Brook University
jcollier@notes.cc.sunysb.edu

Rita Colwell
University of Maryland, College Park
rcolwell@umiacs.umd.edu

Robert Cottingham
Oak Ridge National Laboratory
cottinghamrw@ornl.gov

Aaron Cozen
University of California Santa Cruz
acozen@soe.ucsc.edu

Daniel Cullen
University of Wisconsin-Madison
dcullen@facstaff.wisc.edu

John Cumbers
NASA Ames / Brown University
john.cumbers@nasa.gov

Christina Cuomo
Broad Institute
cuomo@broadinstitute.org

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

Ciara Curtin
Genome Technology
ccurtin@genomeweb.com

Eileen Dalin
Synthetic Genomics
rguyot@syntheticgenomics.com

Chris Daum
DOE Joint Genome Institute
daum1@llnl.gov

Karen Davenport
Los Alamos National Laboratory
kwdavenport@lanl.gov

Evan DeLucia
University of Illinois
delucia@life.uiuc.edu

Li Deng
University of Arizona
ldeng@email.arizona.edu

Shweta Deshpande
DOE Joint Genome Institute
sdeshpande@lbl.gov

Chris Detter
Los Alamos National Laboratory
cdetter@lanl.gov

Ronald DeVries
Utrecht University
r.p.devries@uu.nl

Erika Diaz Almeyda
UC Merced
ediaz-almeyda@ucmerced.edu

Fred Dietrich
Duke University
fred.dietrich@duke.edu

Jayna Ditty
University of St. Thomas
jlditty@stthomas.edu

Jeremy Dodsworth
UNLV
jeremy.dodsworth@unlv.edu

Maria Dominguez
University of Puerto Rico
mgdbello2@gmail.com

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

Sébastien Duplessis
INRA
duplessi@nancy.inra.fr

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

Rob Egan
DOE Joint Genome Institute
rsegan@lbl.gov

Stephanie Eichorst
Los Alamos National Laborator
seichorst@lanl.gov

Jonathan Eisen
UC Davis, DOE Joint Genome Institute
jonathan.eisen@gmail.com

Kevin Eng
DOE Joint Genome Institute
kseng@lbl.gov

Anna Engelbrektson
University of California, Berkeley
aengelbrektson@berkeley.edu

Heinz Falenski
U. C. Merced
hfalenski@ucmerced.edu

Marsha Fenner
DOE Joint Genome Institute
mwfenner@lbl.gov

Alison Fern
DOE Joint Genome Institute
amfern@lbl.gov

Klaus Fiebig
Ontario Genomics Institute
klausfiebig@gmail.com

Christine Foreman
Montana State University
cforeman@montana.edu

Jamie Foster
University of Florida
jfoster@ufl.edu

Cheryl Foust
Oak Ridge National Laboratory
foustcb@ornl.gov

Shari Freyermuth
University of Missouri
freyermuths@missouri.edu

Steven Garan
Lawrence Berkeley National Lab
sgaran@lbl.gov

Maria Ghirardi
NREL
maria.ghirardi@nrel.gov

Filipa Godoy-Vitorino
DOE Joint Genome Institute
fgvitorino@lbl.gov

Barry Goldman
Monsanto
bsgold@monsanto.com

David Goodstein
DOE Joint Genome Institute
DMGoodstein@lbl.gov

Lynne Goodwin
Los Alamos National Laborator
lynneg@lanl.gov

Stephen Goodwin
USDA-ARS / Purdue University
sgoodwin@purdue.edu

Joseph Graber
Department of Energy
joseph.graber@science.doe.gov

Steven Gray
NCSU/Food, Bioprocessing and Nutrition Sciences
srgray@ncsu.edu

Lance Green
Los Alamos National Laborator
ldgreen@lanl.gov

Susan Gregurick
DOE
susan.gregurick@science.doe.gov

Igor Grigoriev
DOE Joint Genome Institute
ivgrigoriev@lbl.gov

Arthur Grossman
Carnegie Institution
arthurg@stanford.edu

Joel Guenther
UC Berkeley
guenthej@berkeley.edu

Lisa Gulino
Agri-Science Queensland
lisa-maree.gulino@deedi.qld.gov.au

Lee Gunter
UT-Battelle/Oak Ridge Natl Lab
gunterle@ornl.gov

Kurt Hafer
Department of Energy
kurt.hafer@oak.doe.gov

Steven Hallam
University of British
Columbiashallam@interchange.ubc.ca

Matthew Hamilton
DOE Joint Genome Institute
mghamilton@lbl.gov

Nancy Hammon
DOE Joint Genome Institute
nmhammon@lbl.gov

James Han
DOE Joint Genome Institute
jkhan@lbl.gov

Shunsheng Han
Los Alamos National Laboratory
han_cliff@lanl.gov

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

Corinne Hausmann
Energy Biosciences Institute
UC Berkeley
corinne.hausmann@gmail.com

Richard Hayes
DOE Joint Genome Institute
rdhayes@lbl.gov

David Hays
DOE Joint Genome Institute
dehays@lbl.gov

Shaomei He
DOE-Joint Genome Institute
she@lbl.gov

Dennis Hedgecock
University of Southern California
dhedge@usc.edu

Karla Heidelberg
University of Southern California
kheidelb@usc.edu

Sabine Heinhorst
The University of Southern Mississippi
sabine.heinhorst@usm.edu

Chris Hemme
University of Oklahoma
hemmecl@ou.edu

Sur Herrera Paredes
UNC
sur00mx@gmail.com

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

Matthias Hess
DOE Joint Genome Institute
mhess@lbl.gov

David Hibbett
Clark University
dhibbett@clarku.edu

Uwe Hilgert
DNALC @ CSHL
hilgert@cshl.edu

Remy Hillekens
NIOO-KNAW
r.hillekens@nioo.knaw.nl

David Hillman
DOE Joint Genome Institute
dwhillman@lbl.gov

Ping Hu
Lawrence Berkeley National Lab
phu@lbl.gov

Phil Hugenholtz
DOE Joint Genome Institute
phugenholtz@lbl.gov

Martin Huh
USC
mhuh@usc.edu

Marcel Huntemann
DOE Joint Genome Institute
mhuntemann@lbl.gov

William Inskeep
Montana State University
binskeep@montana.edu

Natalia Ivanova
DOE Joint Genome Institute
nnivanova@lbl.gov

Javier Izquierdo
Dartmouth College
javier.izquierdo@dartmouth.edu

Sara Jawdy
Oak Ridge National Lab
jawdys@ornl.gov

Chad Jenkins
E&K Scientific
chad@eandkscientific.com

Jerry Jenkins
Hudson Alpha Institute of Biotechnology
jjenkins@hudsonalpha.org

Tomas Johansson
Microbial Ecology, Lund University
tomas.johansson@mbioekol.lu.se

Tracey Kalb-Scherer
Roche Diagnostics
tracey.kalb-scherer@roche.com

Bishoy Kamel
UC Merced
bkamel@ucmerced.edu

Jay Keasling
Joint BioEnergy Institute
Lawrence Berkeley National Lab
University of California, Berkeley
keasling@lbl.gov

Lisa Kegg
DOE Joint Genome Institute
lrkegg@lbl.gov

Martin Keller
Oak Ridge National Laboratory
kellerm@ornl.gov

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

Cheryl Kerfeld
DOE Joint Genome Institute
CKerfeld@lbl.gov

Richard Kerrigan
Sylvan Research
rwk@sylvaninc.com

Madhu Khanna
Univ. of Illinois, Urbana-Champaign
khanna1@illinois.edu

Robin Kodner
University of Washington
rkodner@u.washington.edu

Annegret Kohler
INRA
kohler@nancy.inra.fr

Heidi Kong
National Institutes of Health
kongh@mail.nih.gov

Frank Korzeniewski
DOE Joint Genome Institute
frkorzeniewski@lbl.gov

Heather Koshinsky
Eureka Genomics
heather@eurekagenomics.com

Attendees

Anthony Kosky
DOE Joint Genome Institute
askosky@lbl.gov

Ursula Kuees
University of Gottingen
ukuees@gwdg.de

Victor Kunin
DOE Joint Genome Institute
vkunin@lbl.gov

Alan Kuo
DOE Joint Genome Institute
akuo@lbl.gov

Eiko Kuramae
Netherlands Institute of Ecology
e.kuramae@nioo.knaw.nl

Rainer Kurmayer
Austrian Academy of Sciences
rainer.kurmayer@oeaw.ac.at

Cheryl Kuske
Los Alamos National Laboratory
kuske@lanl.gov

Nikos Kyrpides
DOE Joint Genome Institute
nkyrpides@lbl.gov

Angie Lackey
Roche
angie.lackey@roche.com

Kathleen Lail
DOE Joint Genome Institute
klail@lbl.gov

Miriam Land
Oak Ridge National Laboratory
landml@ornl.gov

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

Peter Larsen
Argonne Natl Laboratory BioSciences
plarsen@anl.gov

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

Sarah Lebeis
University of North Carolina
lebeis@email.unc.edu

Janey Lee
DOE Joint Genome Institute
jlee2@lbl.gov

Tomas Linder
Swedish Agricultural University
ukcommuter@yahoo.com

Erika Lindquist
DOE Joint Genome Institute
ealindquist@lbl.gov

Anna Lipzen
DOE Joint Genome Institute
alipzen@lbl.gov

Wen-Tso Liu
University of Illinois
wtliu@illinois.edu

Connie Lovejoy
IBIS Laval University
connie.lovejoy@bio.ulaval.ca

Todd Lowe
UC Santa Cruz
lowe@soe.ucsc.edu

Steve Lowry
DOE Joint Genome Institute
slowry@lbl.gov

Susan Lucas
DOE Joint Genome Institute
lucas11@llnl.gov

Derek Lundberg
UNC Chapel Hill
derekls@email.unc.edu

Athanasios Lykidis
DOE Joint Genome Institute
alykidis@lbl.gov

Stephanie Malfatti
DOE Joint Genome Institute
malfatti3@llnl.gov

Rex Malmstrom
DOE Joint Genome Institute
rrmalmstrom@lbl.gov

Betty Manfield
Oakridge National Laboratory
manfieldbk@ornl.gov

Vito Mangiardi
DOE Joint Genome Institute
vjmangiardi@lbl.gov

Gerard Manning
Salk Institute
manning@salk.edu

Francis Martin
INRA
fmartin@nancy.inra.fr

Jeffrey Martin
DOE Joint Genome Institute
jamartin@lbl.gov

Joel Martin
DOE Joint Genome Institute
j_martin@lbl.gov

Emma Master
University of Toronto
emma.master@utoronto.ca

Zachariah Mathew
Retired Scientist
nammal6@aol.com

Konstantinos Mavrommatis
DOE Joint Genome Institute
kmavromatis@lbl.gov

Kevin McCluskey
Fungal Genetics Stock Center
mccluskeyk@umkc.edu

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

David Mead
Lucigen
dmead@lucigen.com

Johan Melendez
Johns Hopkins University
jmelend3@jhmi.edu

Xiandong Meng
DOE Joint Genome Institute
xiandongmeng@lbl.gov

Eli Meyer
University of Texas - Austin
elimeyer@mail.utexas.edu

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

Bonnie Millenbaugh
DOE Joint Genome Institute
millenbaugh1@llnl.gov

Thomas Mitchell-Olds
Duke University
tmo1@duke.edu

Martin Mokrejs
Charles University
mmokrejs@iresite.org

Christina Moon
AgResearch Ltd
christina.moon@agresearch.co.nz

Steve Moose
University of Illinois
smoose@illinois.edu

Nancy Moran
University of Arizona
nmoran@email.arizona.edu

Paul Morris
Bowling Green State University
pmorris@bgsu.edu

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

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

Nandita Nath
DOE Joint Genome Institute
nnath@lbl.gov

Rita Nieu
USDA
rita.nieu@ars.usda.gov

Joseph Noel
The Salk Institute
Howard Hughes Medical Institute
noel@salk.edu

Matt Nolan
DOE Joint Genome Institute
mpnolan@lbl.gov

Angela Norbeck
Pacific Northwest National Laboratory
angela.norbeck@pnl.gov

Erin O'Brien
University of Iowa
erin-k-obrien@uiowa.edu

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

Robin Ohm
Utrecht University
r.a.ohm@uu.nl

Jeanine Olsen
University of Groniingen
Centre for Ecol. and Evolutionary Studies
j.l.olsen@rug.nl

Victoria Orphan
Caltech
vorphan@gps.caltech.edu

Giovanni Ortiz
Schafer Corporation
gioscifi@yahoo.com

Jeffrey Osborne
Manchester College
jposborne@manchester.edu

Robert Otillar
DOE Joint Genome Institute
rpotillar@lbl.gov

Diane Ouwerkerk
Agri-Science Queensland
diane.ouwerkerk@deedi.qld.gov.au

William Page
University of Iowa
william-page@uiowa.edu

Jasmyn Pangilinan
DOE Joint Genome Institute
jlpangilinan@lbl.gov

Katherine Pappas
University of Athens
kmpappas@biol.uoa.gr

Ian Paulsen
Macquarie University
ipaulsen@science.mq.edu.au

Yi Peng
DOE Joint Genome Institute
ypeng@lbl.gov

Ze Peng
DOE Joint Genome Institute
zpeng@lbl.gov

Christa Pennacchio
DOE Joint Genome Institute
cppennacchio@lbl.gov

Len Pennacchio
DOE Joint Genome Institute
lapennacchio@lbl.gov

Roger Pennell
Ceres, Inc.
rpennell@ceres.net

Rene Perrier
DOE Joint Genome Institute
raperrier@lbl.gov

Lin Peters
DOE Joint Genome Institute
lgpeters@lbl.gov

Lin Pham
RainDance Technologies
phaml@raindancetech.com

Antonio Pisabarro
Genetics and Microbiology Res. Group
gpisabarro@unavarra.es

Samuel Pitluck
DOE Joint Genome Institute
s_pitluck@lbl.gov

Michael Poulsen
University of Wisconsin, Madison
poulsen@bact.wisc.edu

Amy Powell
Sandia National Laboratories
ajpowel@sandia.gov

Simon Prochnik
DOE Joint Genome Institute
seprochnik@lbl.gov

Teri Rambo Mueller
Roche
teri.mueller@roche.com

Lucía Ramírez
Genetics and Microbiology Res. Group
lramirez@unavarra.es

Kelynne Reed
Austin College
kreed@austincollege.edu

Hans-Joerg Reif
BayerCropscience AG,
hans-joerg.reif@bayercropscience.com

Lee Reilly
DOE Joint Genome Institute
lreilly@lbl.gov

Kathryn Richmond
Great Lakes Bioenergy Res. Center
kerichmond@glbrc.wisc.edu

Robert Riley
DOE Joint Genome Institute
rwriley@lbl.gov

Simon Roberts
DOE Joint Genome Institute
srroberts@lbl.gov

David Robinson
DOE Joint Genome Institute
dsrobinson@lbl.gov

Jorge Rodrigues
University of Texas at Arlington
jorge@uta.edu

Forest Rohwer
San Diego State University
frohwer@gmail.com

Dan Rokhsar
DOE Joint Genome Institute
dsrokhsar@lbl.gov

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

Catherine Ronning
Department of Energy, BER
catherine.ronning@science.doe.gov

Eddy Rubin
DOE Joint Genome Institute
emrubin@lbl.gov

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

Omid Sadeghpour
DOE Joint Genome Institute
osadeghpour@lbl.gov

Asaf Salamov
DOE Joint Genome Institute
aasalamov@lbl.gov

Attendees

Annette Salmeen
DOE Joint Genome Institute
asalmeen@lbl.gov

Erin Sanders
UCLA
erinsl@microbio.ucla.edu

Gustaf Sandh
DOE Joint Genome Institute
gsandh@lbl.gov

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

Steven Savage
Cirrus Partners, LLC
ssavage@cirruspartners.com

Wendy Schackwitz
DOE Joint Genome Institute
wsschackwitz@lbl.gov

Jeremy Schmutz
DOE Joint Genome Institute-HudsonAlpha
jschmutz@hudsonalpha.org

Erin Scully
Pennsylvania State University
eds14@psu.edu

Alexander Sczyrba
DOE Joint Genome Institute
asczyrba@lbl.gov

Harris Shapiro
DOE Joint Genome Institute
hjshapiro@lbl.gov

Louis Sherman
Purdue University
lsherman@purdue.edu

Christine Shewmaker
BluGoose Consulting
blugoose@sbcglobal.net

Weibing Shi
Texas A&M University
wshi@ag.tamu.edu

Prachand Shrestha
Energy Biosciences Institute
prachand@berkeley.edu

David Sims
HudsonAlpha Center for Biotechnology
dsims@hudsonalpha.org

Steven Singer
Lawrence Berkeley National Lab
swsinger@lbl.gov

Kanwar Singh
DOE Joint Genome Institute
ksingh@lbl.gov

Inger Skrede
University of Oslo, Dept of Biology
inger.skrede@bio.uio.no

Steven Slater
Great Lakes Bioenergy Res. Center
scslater@glbrc.wisc.edu

Jason Stajich
University of California, Riverside
jason.stajich@ucr.edu

Wytze Stam
University of Groningen
w.t.stam@rug.nl

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

Jan Stenlid
Swedish University of Agricultural Sciences
jan.stenlid@mykopat.slu.se

Ramunas Stepanauskas
Bigelow Laboratory for Ocean Sciences
rstepanauskas@bigelow.org

Craig Stephens
Santa Clara University
cstephens@scu.edu

Blaire Steven
University of Wyoming
bsteven@uwyo.edu

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

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

Sheng Sun
Duke University Medical Center
sheng.sun@duke.edu

Wesley Swingley
University of California – Merced
wswingley@msn.com

Aijazuddin Syed
DOE Joint Genome Institute
asyed@lbl.gov

Eric Tang
DOE Joint Genome Institute
etang@lbl.gov

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

Tatiana Tatusova
NCBI/NLM/NIH
tatiana@ncbi.nlm.nih.gov

Cameron Thrash
Oregon State University
thrashc@onid.orst.edu

Hope Tice
DOE Joint Genome Institute
tice1@llnl.gov

Tamas Torok
Lawrence Berkeley National Lab
ttorok@lbl.gov

Susannah Tringe
DOE Joint Genome Institute
sgtringe@lbl.gov

Stephan Trong
DOE Joint Genome Institute
trong1@llnl.gov

Miles Trupp
SRI International
trupp@ai.sri.com

Adrian Tsang
Concordia University
tsang@gene.concordia.ca

Anders Tunlid
Lund University
anders.tunlid@mbioekol.lu.se

Gillian Turgeon
Cornell University
bgt1@cornell.edu

Jerry Tuskan
ORNL/ DOE Joint Genome Institute
gtk@ornl.gov

Ludmila Tyler
UC Berkeley
ltyler@berkeley.edu

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

Rytas Vilgalys
Biology Dept., Duke University
fungi@duke.edu

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

Christian Voolstra
KAUST
christian.voolstra@kaust.edu.sa

Hao Wang
University of Georgia
wanghao@uga.edu

Zhong Wang
DOE Joint Genome Institute
zhongwang@lbl.gov

Sarah Watkinson
University of Oxford
sarah.watkinson@plants.ox.ac.uk

Janelle Weaver
UC Santa Cruz
weaver.janelle@gmail.com

Detlef Weigel
Max Planck Institute for Developmental Biology
weigel@weigelworld.org

Willie Wilson
Provasoli-Guillard National Center for Collection of Marine Phytoplankton
wwilson@bigelow.org

Paul Winward
DOE Joint Genome Institute
pwinward@lbl.gov

Dagmar Woebken
Stanford University
dwoebken@stanford.edu

Benjamin Wolfe
Harvard University
bewolfe@gmail.com

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

Alexandra Worden
Monterey Bay Aquarium Res.Institute
azworden@mbari.org

Tanja Woyke
DOE Joint Genome Institute
twoyke@lbl.gov

Crystal Wright
DOE Joint Genome Institute
cawright@lbl.gov

Cindy Wu
Lawrence Berkeley National Lab
chwu@lbl.gov

Dongying Wu
UC Davis Genome Center
dygwu@ucdavis.edu

Guohong Wu
DOE Joint Genome Institute
gwu@lbl.gov

Zhaohui Xu
Bowling Green State University
zxu@bgsu.edu

Xiaohan Yang
Oak Ridge National Laboratory
yangx@ornl.gov

Suzan Yilmaz
DOE Joint Genome Institute
syilmaz@lbl.gov

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

Xiaojing Zhang
Los Alamos National Laboratory
xiaojing.lucy@gmail.com

Xueling Zhao
DOE Joint Genome Institute
xzhao@lbl.gov

Zhiying Zhao
DOE Joint Genome Institute
zyzhao@lbl.gov

Jizhong Zhou
University of Oklahoma
jzhou@ou.edu

Kemin Zhou
DOE Joint Genome Institute
kzhou@lbl.gov

Ruanbao Zhou
South Dakota State University
ruanbao.zhou@sdstate.edu

Author Index

Authors

Authors

Notes