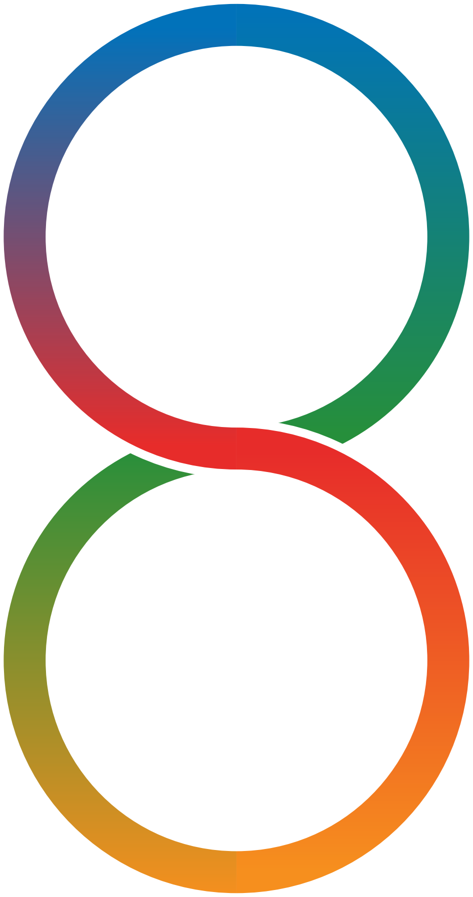
DEPARTMENT OF ENERGY
JOINT GENOME INSTITUTE (DOE JGI)



Genomics of Energy & Environment User Meeting Abstracts

March 26–28, 2013 / Walnut Creek Marriott / Walnut Creek, CA

JGI
DOE JOINT GENOME INSTITUTE

U.S. DEPARTMENT OF
ENERGY
Office of Science

An electronic version of this document is available at:
http://go.usa.gov/g3P

All information current as of March 14, 2013

JGI Contact: Marsha Fenner
DOE Joint Genome Institute
mwfenner@lbl.gov

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

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

Eighth Annual
DOE Joint Genome Institute
User Meeting

Sponsored By

U.S. Department of Energy

Office of Science

March 26-28, 2013

Walnut Creek Marriott

Walnut Creek, California

Contents

Speaker Presentations

Abstracts alphabetical by speaker

Assembly-driven metagenomics of a hypersaline microbial ecosystem

Eric E. Allen (eallen@ucsd.edu)

Marine Biology Research Division, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California; Division of Biological Sciences, University of California at San Diego

Microbial populations inhabiting a natural hypersaline lake ecosystem, Lake Tyrrell, Victoria, Australia, have been characterized using deep metagenomic sampling, iterative *de novo* sequence assembly, and multidimensional phylogenetic binning. Consensus genomes were reconstructed for twelve environmental microbial populations, eleven archaeal and one bacterial, comprising between 0.6-14.1% of the planktonic community. Eight of the twelve genomes recovered represent microbial populations not characterized previously in laboratory culture. The extent of genome assembly achieved has allowed extensive lineage-specific compartmentalization of predicted metagenomic functional capabilities and cellular properties associated with both dominant and less abundant members of the community. The generation of habitat-specific, environmental reference genomes based on metagenomic sequence assembly provides rich opportunities for discovering the genetic architecture of multiple microbial groups constituting a native microbial ecosystem. The degree of connectivity between microbial populations based on metabolic network analysis reveals exquisite functional redundancy in hypersaline microbial communities. This constrained metabolic repertoire may reflect limitations of adaptation to extreme salinity and is consistent with the hypothesis that temporal variability in community membership may be dominated by top-down forcing dynamics (e.g. grazing and viral predation) rather than nutrient availability and assimilation.

Regulation of flowering in *Brachypodium distachyon*

Richard Amasino (amasino@biochem.wisc.edu)

University of Wisconsin, Madison, Wisconsin

The major developmental change in the plant life cycle is the initiation of flowering. Upon the initiation of flowering, many plants do not produce much additional cellulosic biomass for the remainder of the growing season. The initiation of flowering is often a response to sensing seasonal change—for example sensing changes in day-length and temperature. We are using a genetic approach to identify genes that are involved in the timing of flowering in the model grass *Brachypodium distachyon*, with a focus on vernalization—the acquisition of competence to flower after prolonged exposure to winter cold. Although work from our group and others has revealed much about the molecular nature of vernalization in the crucifer *Arabidopsis thaliana*, grasses have a vernalization system that evolved independently of that in *Arabidopsis thaliana*. In fact, the vernalization systems in different groups of plants result from convergent evolution because when flowering plants were diverging there was not an advantage to possessing vernalization systems—the Earth was much warmer and the landmasses were in different locations than at present. Our study of biomass traits in *Brachypodium distachyon* is supported by the Great Lakes Bioenergy Research Center (www.glbrc.org/).

Network as discovery instrument: A quick-start guide

Greg Bell (grbell@es.net)

DOE ESnet, Lawrence Berkeley National Laboratory, Scientific Networking Division, Berkeley, CA

Scientific progress should not be constrained by the physical location of instruments, people, computational resources, or data. Advanced research networks are now making steady progress at achieving that vision. In this talk, ESnet's Greg Bell will attempt to raise your expectations for networks; explain how they can accelerate scientific workflows; and provide guidance about getting maximum benefit from ESnet, Internet2, or whichever research network serves your home institution.

The Isthmus of Panama — A natural laboratory for the genomics of climate resilience

Eldredge Bermingham (Bermingham@si.edu)

Smithsonian Tropical Research Institute, Washington, D.C.

The Isthmus of Panama joined the continents of North and South America roughly 3 million years ago, and separated the modern Caribbean Sea from the Eastern Pacific Ocean. This event established an interchange of terrestrial biodiversity of epic proportions between the continents, and initiated the genomic divergence of marine organisms separated by the isthmus. Moreover, the narrow isthmus spans a two-fold rainfall gradient in the 300,000-hectare Panama Canal Watershed, and is impacted by El Niño and La Niña. The environmental attributes of Panama, coupled to the 100-year presence of the Smithsonian Tropical Research Institute on the Isthmus, create a remarkable opportunity to use genomic approaches to understand diversity and climate resilience in organisms ranging from the coral holobiont to the mosaic genomes of rain-forest trees.

Single-cell genomics of environmental bacteria

Paul Blainey (pblainey@broadinstitute.org)

Broad Institute

Succession of microbial phylogeny and function during plant litter decomposition

Eoin L. Brodie (elbrodie@lbl.gov) and Mari Nyyssonen

Ecology Department, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Trait-based models that connect microbial composition to ecosystem functioning typically represent microorganisms as functional guilds or groups of organisms sharing similar trait combinations. Key to this approach is the need to constrain trait combinations to avoid the emergence of so called ‘Darwinian demons’ – an organism that maximizes all fitness traits. Advances in environmental metagenomics now permit reconstruction of complete genomes from complex microbial communities, potentially enabling trait distributions and trait combinations to be defined or at least constrained.

However, this approach is largely dependent on *in silico* gene predictions and is facilitated by characterized microbial isolates of limited diversity and environmental distribution – which may result in an underestimation of functional diversity and redundancy. Parallel function-based mining of metagenomic libraries can provide direct linkages between genes and metabolic traits and thus bridge the gap between sequence data generation and functional predictions from environmental genomic datasets. We have developed high-throughput screening assays for function-based characterization of activities related to plant polymer decomposition contained in environmental metagenomes. The multiplexed assays are based on fluorogenic and chromogenic substrates and combine automated liquid handling and a genetically modified expression host to enable the daily screening of 12,160 fosmid clones against 13 substrates in more than 170,000 reactions. To illustrate the utility of the developed assays, we analyzed leaf litter metagenomic fosmid libraries sampled during litter decomposition. Positive clones were sequenced using the MiSeq platform for identification of genes conferring the observed activities and associated traits. Merging these data with shotgun metagenomic sequence data allows the genomic context of these traits to be determined and the linkage of additional functional traits.

Strange tails in viral discovery

Joseph DeRisi (joe@derisilab.ucsf.edu)

Howard Hughes Medical Institute, University of California at San Francisco, San Francisco, California

Synthetic metagenomics: Converting digital information back to biology

Samuel Deutsch (sdeutsch@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

Next-generation sequencing technologies have enabled single genomes as well as complex environmental samples (metagenomes) to be sequenced on a routine basis. Bioinformatics analysis of the resulting sequencing data reveals a continually expanding catalogue of predicted proteins (> 50 million), which cover the full spectrum of known pathways and functional activities. Converting sequencing data retrieved from databases (digital information) into biochemical molecules that can be functionally characterized remains challenging for a number of reasons. One way to overcome the 'information gap' is through synthetic biology methods that allow genes and pathways to be synthesized in a template independent manner. Over the past 2 years we have developed an automated DNA synthesis pipeline at JGI, and have produced several megabases of synthetic DNA for internal as well as user projects. The projects include high-throughput characterization individual enzymes, generation of combinatorial libraries for synthetic pathways, and genome-scale engineering. In addition, we are working on informatics solutions for DNA synthesis design and implementation, an area that is becoming increasingly important as the field develops.

Biodiversity monitoring using NGS approaches on unusual substrates

Tom Gilbert (mtpgilbert@gmail.com)

University of Copenhagen, Copenhagen, Denmark

The advent of NGS has not just revolutionised genomic studies, but is playing an increasingly important role in eukaryotic biodiversity monitoring. In particular, the coupling of NGS platforms with taxa generic primer sets represents an attractive alternative to conventional cloning/Sanger sequencing, that enables incredibly deep sequencing of DNA extracted from complex substrates. Although extraction of DNA from such substrates is often tricky, if successful these can often yield genetic material derived from a wide range of plants and animals, providing a direct insight into the communities that relate to the substrate. In this presentation I demonstrate the contribution of some surprising substrates to biodiversity assessments, enabling the targetting of a wide range of questions including how ecosystems have changed through time, and what animals are present in environments that are traditionally hard to survey. Furthermore, I briefly outline some of the challenges in the generation and analysis of such data, and suggest future avenues to be developed.

Genetic regulation of grass biomass accumulation and biological conversion quality

Sam Hazen (hazen@bio.umass.edu)

University of Massachusetts, Amherst, Massachusetts

TARA OCEANS: A global analysis of oceanic plankton ecosystems

Eric Karsenti (karsenti@embl.de)

EMBO Heidelberg, Heidelberg, Germany

TARA OCEANS is a research project that started with a round-the-world expedition aboard the 110-foot schooner TARA in September 2009 and ended in March 2012. The goal of the expedition was to sample as quantitatively as possible pelagic planktonic organisms from viruses to zooplankton, ranging in size over 6 orders of magnitude. To be useful for ecological studies, the sampling was planned to provide a large number of environmental parameters associated with biological sampling, from most oceans including North Atlantic, the Mediterranean Sea, Red Sea, Indian Ocean, South Atlantic and Antarctica, South and North Pacific, Gulf Stream and Sargasso Sea. In addition to global sampling, the localization of each station was carefully chosen as a function of currentology and local ocean state, within the context of specific water masses derived from satellite imagery. This was done in order to be able to project the potential evolution of the local ecosystems in time after having established correlations between ecosystem biological compositions and their physical environment, from the full round-the-world data set. Oceanographic parameters are all validated and stored at Pangaea. Organisms and genes are analyzed using a combination of automated imaging methods adapted to each size class and metagenomics analysis. Because the expedition just came back one year ago, the consortium is still in the process of collecting and organizing data sets. However, some preliminary results start to emerge, in particular about the level of biodiversity that might exist across the size range we have examined, the level of gene content knowledge we have concerning these organisms, the relationship between some environmental parameters like oxygen concentration and organisms compositions, the co-occurence of various types of organisms, etc. In the seminar I will present the rationale of the expedition and analysis strategy, the methods used as well as some preliminary results. I will also raise issues about the complexity of the analysis of these kinds of ecosystems as well as data access and connectivity.

Systems genetics and genomics of woody biomass production in *Eucalyptus*, a global fibre crop

Alexander A. Myburg* (zander.myburg@fabi.up.ac.za),[1] Eshchar Mizrachi,[1] Dario Grattapaglia,[2,3] Gerald A. Tuskan,[4,5] The *Eucalyptus* Genome Network (EUCAGEN)[6]

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa; [2]Plant Genetics Laboratory, EMBRAPA Genetic Resources and Biotechnology, Brasilia, Brazil; [3]Genomic Sciences Program, Universidade Católica de Brasília, Brasília, Brazil; [4]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]DOE Joint Genome Institute, Walnut Creek, California; [6]www.eucagen.org

Their wide adaptability, fast growth and superior fibre properties have driven the global increase in cultivation of *Eucalyptus* tree species and hybrids for plantation forestry (now >20 Mha world-wide). Woody biomass from eucalypt plantations is processed for pulp, paper and chemical cellulose, and has been identified as a potential feedstock for bioenergy and biomaterials. The DOE-JGI has completed a draft genome sequence for *E. grandis,* a widely grown subtropical forestry species, serving as a reference for the genus and the first representative genome for the basal Rosid order Myrtales. The draft genome assembly (http://www.phytozome.net/eucalyptus.php) was produced from 7.7 million Sanger reads (8X coverage) and comprises 691 Mb, of which 605 Mbp is assembled in 11 mapped chromosome scaffolds (N/L50: 5/53.9 Mb). A total of 36,376 loci are predicted to contain protein-coding transcripts including key regulatory and biosynthetic gene families underlying the distinctive growth, development and secondary metabolism of eucalypts. To further investigate the genetic architecture of woody biomass traits in *Eucalyptus*, we have sequenced (RNA-Seq, 20 million PE50) and profiled the developing xylem transcriptomes of 280 interspecific (F2) backcross progeny of an *E. grandis* x *E. urophylla* F1 hybrid clone. The xylem transcript abundance profiles of 30,471 genes were used to identify more than 27,000 expression quantitative trait loci (eQTLs) with cis-acting (29%) and/or trans-acting (65%) effects on transcript level variation in F2 progeny. Trans-eQTLs frequently occur clustered in "hotspots" in the F1 hybrid genome marking the location of major regulatory loci affecting the expression of hundreds of genes elsewhere in the genome. Together with cell wall chemistry traits measured in the same trees, eQTL and population-wide correlation (gene-gene and gene-trait) data allow a systems genetics approach to unravel the regulation of woody biomass production in mature, field grown *Eucalyptus* trees. Furthermore, we show how recently developed genotyping resources (DArT and a high density 60K SNP EuCHIP) for eucalypt species have been used for QTL/gene discovery and molecular breeding by genome-wide prediction of complex growth and wood property traits.

Automated strain engineering at Amyris

Jack Newman (Newman@amyris.com)

Amyris Inc.

Uncovering transcriptional circuits in Arabidopsis by functional genomics

Jose Pruneda-Paz (jprunedapaz@ucsd.edu)

University of California at San Diego, San Diego, California

Extensive transcriptional networks control every biological process in plants. Myriad environmental and endogenous cues regulate specific gene expression profiles ultimately determining overall plant fitness and performance. While multiple mechanisms control transcript abundance, specific transcription factors (TFs) that regulate the expression of any given Arabidopsis gene are still widely unknown. The redundant nature of transcriptional circuits and TF function significantly limit discovery approaches. To circumvent this problem, we implemented TF-centered yeast one-hybrid (Y1H) screens as an alternative means to uncover direct regulators of a key Arabidopsis clock component (CCA1). The strategy, using an initial 200 TF-collection, was instrumental in identifying a novel clock component (CHE) that directly regulates *CCA1*. Based on this success we aimed to generate a versatile research tool that could be widely used for the discovery of TFs. Here, we present the construction of the most comprehensive fully sequence-validated collection of Arabidopsis TFs. This 1956-clone collection includes ~80% of all TFs predicted by most TF-specific databases and provides a significant increase in TF coverage compared to similar, previously generated public resources. Furthermore, the collection is completely homogeneous, as all clones were generated following the same cloning strategy, and compatible with recombination based cloning, so TF-coding sequences can be easily transferred to different vectors for downstream applications. This capability allowed us to generate a collection of TF-constructs suitable for Y1H screens. Using this novel collection, we implemented an improved procedure to perform high-throughput Y1H automated screens that was used to identify novel regulators of *CCA1*. In summary, the resources presented here provide a comprehensive toolset for the discovery of direct transcriptional regulators for any Arabidopsis gene.

The genomic encyclopedia for archaea and bacteria-root nodule bacteria (GEBA-RNB) community sequencing project: A comprehensive resource of microsymbiont genomes

Wayne G. Reeve* (W.Reeve@murdoch.edu.au),[1] Lambert Brau,[2] Natalia Ivanova,[3] and Nikos Kyrpidis[3]

[1]Centre for *Rhizobium* Studies, School of Biological Sciences and Biotechnology, Murdoch University, Murdoch, Australia; [2]Deakin University, Melbourne Burwood Campus, Burwood, Victoria, Australia; [3]DOE Joint Genome Institute, Walnut Creek, California

Genome sequencing has revolutionized many aspects in microbiology particularly with regards to pathogenesis, energy production, bioremediation, global nutrient cycles, and the origins, evolution, and diversity of life. The currently available microbe genome sequences show a highly biased phylogenetic distribution compared to the extent of microbial diversity known today. This bias has resulted in a major gap in our knowledge of microbial genome complexity and our understanding of the evolution, physiology, and metabolic capacity of microbes. There is still a need for a large-scale systematic effort to sequence microbial genomes to fill in genomic gaps in the tree of life. To meet this need, the Genomic Encyclopedia of Bacteria and Archaea (GEBA) program was initiated. One branch of the GEBA project has concentrated on establishing the genomes of 100 Root Nodule Bacteria (GEBA-RNB) that are legume microsymbionts. Currently, 90% of the genomes have been completed and deposited into various databases including NCBI, GOLD and IMG-GEBA. The 100 RNB strains have commercial, genetic and ecological importance and comprehensive metadata exists for these strains. These strains have been sourced from the rhizobial community with the aim to study biogeographical effects on species evolution, identify unique genetic attributes that may provide superior adaptation and identify host specificity and nitrogen fixation determinants. The GEBA-RNB project involves the US Joint Genome Institute (US) and an international consortium of 30 collaborators from 15 countries. An overview of the project and some recent findings will be presented.

Genetics and genomics fuel the development of the energy crop *Jatropha curcas*

Bob Schmidt (bschmidt@sgbiofuels.com)

SG Biofuels; University of California, San Diego

Part mining for synthetic biology

Chris Voigt (cavoigt@gmail.com)

Massachusetts Institute of Technology, Cambridge, Massachusetts

The challenges and opportunities for extending plant genomics to climate

David Weston (westondj@ornl.gov)

Oak Ridge National Laboratory, Oak Ridge, Tennessee

A glimpse into the coding potential of microbial dark matter

Tanja Woyke (twoyke@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

Poster Presentations

Posters alphabetical by first author. *Presenting author

Screening the tropical fungal biodiversity of Vietnam for biomass modifying enzymes, with secretome and transcriptome analyses

George E. Anasontzis* (George.anasontzis@chalmers.se)**,**[1,3] Dang Tat Thanh,[2] Nguyen Thanh Thuy,[2] Dinh Thi My Hang,[2] Vu Nguyen Thanh,[2] and Lisbeth Olsson[1,3]

[1]Department of Chemical and Biological Engineering, Industrial Biotechnology, Chalmers University of Technology, Gothenburg, Sweden; [2]Microbiology Department, Food Industries, Research Institute, Thanh Xuan, Hanoi, Vietnam; [3]Wallenberg Wood Science Center, Chalmers, Gothenburg, Sweden

In the bio-based economy concept, the current hydrocarbon fuels and non-biodegradable plastics will be replaced by new products which will derive from natural and renewable resources. The synthesis of such biofuels and biochemicals is still challenged by the difficulties to cost efficiently degrade lignocellulosic materials to fermentable sugars or to isolate the intact polymers. Biomass degrading and modifying enzymes play an integral role both in the separation of the polymers from the wood network, as well as in subsequent modifications, prior to further product development. The type of application usually defines the conditions where the reactions should take place. Thus, novel enzymes with variable combined properties, such as different thermotolerance, pH range of activity, substrate specificity and solvent tolerance, still need to be discovered and developed to achieve the highest possible efficiency in each occasion. We took advantage of the rapidly evolving and high biodiversity of the tropics and have been screening various isolates for their cellulases and hemicellulases activities. Promising strains were then cultivated in bioreactors with different carbon sources, such as wheat bran, spruce and avicel and their biomass degrading capacity was analysed through cross species protein identification of their secretome with TMT. Information on the genes involved in the different stages of the fermentation and the carbon source will be acquired with next generation sequencing of the total transcriptome. Interesting transcripts will then be used to heterologously clone and express the respective genes and identify their role in the degradation process.

KBase: An integrated knowledgebase for predictive biology and environmental research

Adam Arkin[1] (aparkin@lbl.gov), Robert Cottingham,[2] Sergei Maslov,[3] Rick Stevens,[4] Dylan Chivian,[1] Parmavir Dehal,[1] Christopher Henry,[4] Folker Meyer,[4] Jennifer Salazar,[4] Doreen Ware,[5] David Weston,[2] **Brian Davison***,[2] and Elizabeth M. Glass[4]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Brookhaven National Laboratory, Upton, New York; [4]Argonne National Laboratory, Argonne, Illinois; and [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

The new Systems Biology Knowledge base, or KBase is integrating commonly used core tools and their associated data, and building new capabilities on top of the combined data. New functionality allows users to visualize data, create powerful models or design experiments based on KBase-generated suggestions. Although the integration of different data types will itself be a major offering to users, the project is about much more than data unification. KBase is distinguished from a database or existing biological tools by its focus on interpreting missing information necessary for predictive modeling, on aiding experimental design to test model-based hypotheses, and by delivering quality-controlled data. The project leverages the power of cloud computing and high-performance computing resources across the DOE system of labs to handle the anticipated rapid growth in data volumes and computing requirements of the KBase. KBase is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It is a community-driven, extensible and scalable open-source software framework, and application system. KBase offers free and open access to data, models and simulations, enabling scientists and researchers to build new knowledge, test hypotheses, design experiments, and share their findings.

Comparative metabolite profiling of cyanobacteria

Richard Baran* (RBaran@lbl.gov),[1] Benjamin P. Bowen,[1] Nicholas Jose,[1,2] Vamsi Moparthi,[2] Cheryl A. Kerfeld,[2,3] Ferran Garcia-Pichel,[4] Muriel Gugger,[5] and Trent R. Northen[1]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of California at Berkeley, Berkeley, California; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]School of Life Sciences, Arizona State University, Tempe, Arizona; [5]Institut Pasteur, The Pasteur Culture Collection of Cyanobacteria, Paris, France

Cyanobacteria play an important role in global carbon cycling as primary producers in diverse environments. We have previously used untargeted mass spectrometry-based metabolite profiling to identify a number of unexpected metabolites and unexpected metabolic capabilities (uptake or release of metabolites) in a model unicellular cyanobacterium *Synechococcus* sp. PCC 7002. These metabolites and metabolic capabilities were classified as unexpected because their presence was not predicted from the genome’s annotation or, in multiple cases, also not accounted for in databases of metabolism (MetaCyc or KEGG). In order to link these metabolic capabilities to specific genes, we performed metabolite profiling on ten different strains of cyanobacteria which were sequenced as part of the CyanoGEBA project. We detected the presence of metabolites of interest (e.g. histidine betaine and derivatives, diverse oligosaccharides) only in subsets of profiled cyanobacteria. These results along with available genome sequences serve as the basis for the identification of corresponding candidate biosynethetic genes. In addition to metabolite profiling we also performed stable isotope probing experiments to measure the degree of turnover of intracellular metabolites. We found that some metabolites are turned over in correlation with biomass growth while others are turned over extensively over shorter periods of time. These results provide an overall view of the carbon flow in the cell as well as point to suitable starting points for engineering heterologous pathways for biotechnological applications.

Metagenomics for plant litter deconstruction in natural and perturbed environments

R. Berlemont* (rberlemo@uci.edu),[1,2] S.D. Allison,[1,2] J.B.H. Martiny,[2] E. Brodie,[3] and A.C. Martiny[1,2]

[1]Department of Earth System Science, University of California at Irvine, Irvine, California; [2]Department of Ecology and Evolutionary Biology, University of California at Irvine, Irvine, California; [3]Department of Environmental Science Policy & Management, University of California at Berkeley, Berkeley, California

Plant litter degradation by microbial consortia is a global process occurring in all terrestrial ecosystems and is a key step in the global C-cycling. In many land ecosystems, climate change (e.g. reduced water availability) and nitrogen deposition are known to alter biological communities and environmental processes. These two drivers are known to affect the leaf-litter decomposition rate. Drought significantly reduces the plant litter decomposition whereas nitrogen increases the deconstruction rate.[1] Comparative genomic analysis allows discriminating potential degraders from other microbes regarding the genomic content for glycoside hydrolases (e.g. cellulases). However, most of the completely sequenced microorganisms are unable to degrade the cellulose.[2] In addition, little is known about the structure of environmental microbial communities inhabiting the litter. Thus, in order to understand how global changes affect the plant polymer processing in the litter, insights into the microbial community structure are required: Which microorganisms inhabit the plant litter? How environmental perturbations will affect the community structure and the distribution of genes for cellulose deconstruction? To address these questions, we conducted a metagenomics study on the microbial community from plant litter derived from natural and perturbed environments. Plant litter from Loma Ridge experimental fields was sampled once every season during two years. DNA was extracted from the different following treatments: control, added nitrogen and reduced water, sequenced by high-throughpout Illumina sequencing (110 Gbp). According to our analysis, fluctuations in water availability have a strong effect on the abundance and distribution of microorganisms. Control (m.a.p. 325mm/y) and artificially “half-reduced water” plots display similar microbial populations (*P*>0.8). On the opposite, chronic nitrogen deposition produces a distinct, less fluctuating, microbial population. However, during the dry season, water depletion is the most important factor affecting the microbial community structure, the abundance and the distribution of features involved in plant litter deconstruction. In addition, the analysis of the distribution of microbial taxa regarding the treatments, during the two years of the experiment, highlights resilient and sensitive microbial groups. More generally, our analysis reveals how environmental changes can impact microbial communities involved in the C-cycling. This is important for understanding and predicting accurately the impact of environmental changes on global processes.

[1] **Allison, S. D**., Y. Lu, C. Weihe, M. L. Goulden, **A. C. Martiny**, K. K. Treseder, and **J. B. H. Martiny**. 2012. Microbial abundance and composition influence litter decomposition response to environmental change. Ecology in press.

[2] **R. Berlemont** and **A. C. Martiny**. 2013. Phylogenetic distribution of potential cellulases in bacteria. Appl. Environ. Microbiol. 79:5:1545-54. 2013.

Metabolic engineering of *Clostridium thermocellum* for biofuel production from cellulosic substrates

Ranjita Biswas,[1,2] Nannan Jiang,[1,2] Lee R. Lynd,[2,3] and Adam Guss* (gussam@ornl.gov)[1,2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]BioEnergy Science Center, Oak Ridge, Tennessee; [3]Dartmouth College, Hanover, New Hampshire

Clostridium thermocellum is a leading candidate organism for implementing a consolidated bioprocessing (CBP) strategy due to its native ability to rapidly consume cellulose and its existing ethanol production pathway. Since substrate costs are a significant fraction of final costs, high product yield is needed for commercial viability. *C. thermocellum* converts cellulose and cellobiose to lactate, formate, acetate, hydrogen, ethanol, amino acids, and other products. However, the mechanism for flux distribution at the various metabolic branch points is not well understood. Furthermore, while pyruvate kinase typically converts phosphoenolpyruvate (PEP) to pyruvate during glycolysis, it is absent from the *C. thermocellum* genome sequence. The lack of pyruvate kinase leads to the hypothesis that *C. thermocellum* utilizes an unusual glycolytic pathway in which PEP carboxykinase converts PEP to oxaloacetate, malate dehydrogenase converts oxaloacetate to malate, and malic enzyme converts malate to pyruvate. To better understand central metabolism and to increase ethanol yield, we have constructed *C. thermocellum* deletion and overexpression mutants to constrain flux of carbon and electrons through glycolysis. Multiple approaches are being undertaken, including alteration of alcohol dehydrogenase cofactor specificity, deletion of genes involved in production of H_2, acetate, lactate, and formate, and strain evolution. The resulting strains produce substantially more ethanol than the wild type and are being further modified by both rational and random approaches. Phenotypic and genotypic characterization of these strains gives insight into central metabolism of *C. thermocellum* and suggests future paths for the engineering of more efficient biofuel production from lignocellulosic biomass.

A Functional Encyclopedia of Bacteria and Archaea

M.J. Blow* (mjblow@lbl.gov),[1] A.M. Deutschbauer,[2] C.A. Hoover,[1] M.N. Price,[2] K.M. Wetmore,[2] A.P. Arkin,[2] and J. Bristow[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Bacteria and Archaea exhibit a huge diversity of metabolic capabilities with fundamental importance in the environment, and potential applications in biotechnology. However, the genetic bases of these capabilities remain unclear due largely to an absence of technologies that link DNA sequence to molecular function. Indeed, in the typical bacterial genome over a quarter of genes are of unknown function. To address this challenge, we are developing a novel platform for gene function annotation by combining high throughput experimental and DNA sequencing approaches, and will use this platform to annotate gene function in 50 diverse microbes. Our strategy is to first determine the phenotypic capabilities of each microbe using high-throughput growth assays under ~300 diverse growth conditions. To understand the genetic bases of these capabilities, we use transposon mutagenesis to generate a complex mutant population for each microbe. Within a mutant population, individual cells harbor a single transposon insertion, but collectively the population contains at least one inactivating mutation for every gene in the genome. By growing mutant populations under a large number of relevant growth conditions, and monitoring changes in the abundance of individual mutant strains by DNA sequencing, we will determine the importance of each gene for survival under every growth condition, and thus derive a set of high-level functional annotations for every gene in the genome. In a proof of principle study in a single microbe, this approach yielded novel functional annotations for over 100 (25%) genes with previously unknown function. By extending this approach to 50 diverse organisms we expect to uncover hundreds of novel gene functions, and to refine the annotations of thousands of other genes. Data from this study will constitute a proof of principle 'Functional Encyclopedia of Bacteria and Archaea' and will be incorporated into the DOE KnowledgeBase system to enable improved modeling of metabolic pathways, and to functionally annotate existing and future microbial genome sequences.

Environmental drivers of microbial ecology in the dry valleys of Antarctica

Eric M. Bottos,[1] Charles K. Lee* (cklee@waikato.ac.nz),[1] Daniel Laughlin,[1] S. Craig Cary[1, 2]

[1]International Centre for Terrestrial Antarctic Research, University of Waikato, Hamilton, New Zealand; [2]College of Earth, Ocean, and Environment, University of Delaware, Lewes, Delaware

The McMurdo Dry Valleys of Antarctica harbor a unique ecosystem for which abiotic factors influencing microbial communities predominate and can be readily identified. The trophic structure within the Dry Valleys is very well characterized and likely represents the simplest carbon cycle on the planet. The New Zealand Terrestrial Antarctic Biocomplexity Survey aims to describe the composition of microbial communities in the Dry Valleys and to elucidate how their structure is driven by environmental factors at the landscape level. Over two field seasons, more than 600 soil samples were collected from a 250-km^2 area, encompassing the Miers, Marshall, and Garwood Valleys. Using DNA fingerprinting techniques and high-throughput DNA sequencing of 16S rRNA gene, we characterized the bacterial communities in more than 450 samples. Bacterial community structures were found to vary significantly ($p<0.001$) with respect to many of the landscape and physicochemical variables measured, but most strongly with variation in elevation and soil moisture. From our comprehensive collection of biological and environmental data, we have constructed a structural equation model (SEM) to describe how environmental drivers interact with microbial diversity, and this model is being used to identify *key spots* where abiotic conditions are likely to have triggered specific metabolic adaptataions in local microbial communities. Through the JGI Community Sequencing Program, we are in the process to resolve how environmental drivers interact with microbial functions in this unusual and highly constrained ecosystem by performing metagenomic analyses of such *key spots*. Our findings will shed light on the abiotic drivers of microbial ecology and function and provide a critical set of information on ultraoligotrophic ecosystems for the Earth Microbiome Project.

Easy Terminal Alternative (ETA): An intuitive Web portal to simplify computational analysis, visualisation, and collaboration

Alexander E. Boyd* (aeboyd@lbl.gov),[1,2] Christopher M. Sullivan,[1,2] and Brett M. Tyler[1]

[1]Center for Genome Research and Biocomputing, [2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon

ETA is an ambitious project that is focused on delivering an easy-to-use web-portal for data analysis for the novice user, as well as a framework in which developers can rapidly develop new bioinformatic tools. Any command line tool can be dynamically represented by a "wrapper" in a straightforward web form. A wrapper contains information such as: inputs, outputs, program path, and environment variables needed by the target program. Wrappers can be made by anyone and then easily shared or put into a centralized community-driven repository. Syncing wrappers across multiple instances of ETA is supported. Wrappers are tied together with a drag-and-drop interface to create a workflow. ETA automatically handles concurrency issues and workflow monitoring that makes for quick recovery. Pausing, restarting, and modifying inputs on processing workflows is supported. Email and/or SMS messages can be sent if any failure occurs that can not be automatically recovered. ETA is contained in a java servlet that can be deployed to many different web services. It utilizes MySQL to store all required information and can use any authentication system or grid engine by development of a plugin.

The characterization of endophytic rhizobacteria isolates from *Arabidopsis thaliana*

Natalie Breakfield* (nbreakfield@gmail.com),[1] Sur Herrera Paredes,[1] Derek Lundberg,[1] Sarah Lebeis,[1] Scott Clingenpeel,[2] Tijana Glavina del Rio,[2] Julien Tremblay,[2] Susannah Tringe,[2] and Jeff Dangl[2]

[1]University of North Carolina, Chapel Hill, North Carolina; [2]DOE Joint Genome Institute, Walnut Creek, California

There is a direct link between plant productivity and the plant microbiome, both in the narrow region of soil directly influenced by root exudates (the rhizosphere) and within the root itself (endophytic compartment). Previous studies in our lab showed that Actinobacteria and Proteobacteria are enriched in the endophytic compartment in *Arabidopsis thaliana*. We hypothesize that members of these bacterial phyla are providing benefit to the plant, presumably through increased nutrient uptake or pathogen defense. In this project, we are characterizing endophyte-isolated bacteria under standard and phosphate nutrient stress conditions to identify plant growth promoting (PGP) activities. The Dangl lab has already isolated ~400 bacteria from endophytic compartments from two soils, and ~85 of these bacteria match the ~170 OTUs that we identified as being enriched in the endophytic compartment compared to bulk soil. All of these isolates are being sequenced and annotated at JGI. We have focused our initial screens for plant growth promotion on two genera of Proteobacteria that often have known PGP activities across broad host phyla, namely the α-Proteobacteria *Rhizobium* and the γ–Proteobacteria *Pseudomonas*. To date, 8 isolates show statistically significant PGP activity, with 4 exhibiting strong PGP activity. In addition to mono-association assays, we are also assaying combinations of bacteria with and without PGP activities on plates and in our synthetic soil substrate. Initial results indicate that the presence of a specific Actinobacteria with no PGP activity seems to affect the number of a Proteobacteria that are found in the endophytic compartment. Future studies include sequencing transcripts from both plants and bacteria in association to identify the pathways important in the endophytic association and PGP activities, and using fluorescently tagged isolates to visualize the re-colonization of simple communities over root developmental time.

The KBase architecture and infrastructure design

Thomas Brettin* (brettints@ornl.gov),[1] Robert Olson,[2] Ross Overbeek,[2] Terry Disz,[2] Bruce Parello,[2] Shiran Pasternak,[5] James Gurtowski,[5] Folker Meyer,[2] Michael Galloway,[1] Steve Moulton,[1] Dan Olson,[2] Shane Canon,[3] Shreyas Cholia,[3] Dantong Yu,[4] Shinjae Yoo,[4] Pavel Novichkov,[3] Daniel Quest,[1] Narayan Desai,[2] Jared Wilkening,[2] Miriam Land,[1] Scott Deviod,[2] Adam Arkin,[3] Robert Cottingham,[1] Sergei Maslov,[4] and Rick Stevens[2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Argonne National Laboratory, Argonne, Illinois; [3]Lawrence Berkeley National Laboratory, Berkeley, California; [4]Brookhaven National Laboratory, Upton, New York; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

The Systems Biology Knowledgebase (KBase) has two central goals:

1. The scientific goal is to produce predictive models, reference datasets and analytical tools for research relating to bioenergy, the carbon cycle, and the study of subsurface microbial communities.

2. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance and use of predictive models and methods in the study of microbes, microbial communities and plants.

The driving objectives of the KBase architecture and infrastructure design focus on creating an unprecedented user experience. The integrated software and hardware infrastructure supporting the user experience comprises a continuously expanding collection of software and services. These are hosted on a physical infrastructure consisting of high speed wide area networking, cloud computing resources, and state of the art cluster computing resources.

Forest harvesting alters the genetic potential for lignocellulose decomposition in soil communities

Erick Cardenas Poire* (carden24@mail.ubc.ca), Kendra Mitchell, Melanie Scofield, Steven Hallam, and William H. Mohn

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

Forest are complex systems that support biodiversity, capture and store carbon, and contribute to Canada's Economy. It is important to understand the effects of forest management practices on microbial metabolism in order to manage forest resources in a way that is sustainable and minimizes global warming. We used metagenomic analysis to investigate the effect of organic matter (OM) removal on lignocellulose degradation potential in forest soil communities. Our study site, located at O'Connor Lake, British Columbia, is part of the Long Term Soil Productivity Study (LTSP), which comprises many forest sites across North America. We sampled our site 13 years after harvesting, which involved treatments with levels of OM removal, and replanting with Lodgepole Pine. DNA was extracted from triplicate samples from organic and mineral horizons and directly sequenced using Illumina Hiseq technology which provided us around 150 million paired end sequences per sample. After quality control, sequences were compared against two databases of enzymes involved in Carbohydrate (CAZy), and Lignin (FOLy) degradation via blastx using a 10E-5 evalue threshold. After normalization, we detected statistically significant decrease in the relative abundance of CAZy and FOLy enzymes as OM removal increase. The difference was significant even and the least severe level of OM removal and was observed in both mineral and organic horizons. The majority of variation (>70%) in CAZy and FOLy gene profiles was significantly ($p<0.001$) explained by the OM removal treatment and the soil horizon, according to a constrained analysis of principal coordinates ordination. Variation in CAZy and FOLy genes was significantly explained by the OM removal and soil horizon. Soil horizon was the principal factor explaining CAZy gene variation (explained 61.6%) whereas OM removal was the main factor explaining FOLy gene variation (explained 41.1%). Analysis of variation was done by PERMANOVA tests, and results were significant at $p< 0.01$. The strong influence of soil horizon on CAZy gene profiles may correspond to distinct bacterial communities in the two horizons, while the weaker influence of horizon on FOLy gene profiles may correspond to similar fungal communities in the two horizons. These results suggest that forest harvesting may have long-term effects on the capacity of soil communities to decompose lignocellulose even when no effect on tree productivity has been yet observed.

Sequence consensus algorithms and hierarchical genome assembly process for effective *de novo* assembly with SMRT® sequencing

Jason Chin* (jchin@pacificbiosciences.com), Patrick Marks, David H. Alexander, Aaron Klammer, Michael Brown, and Cheryl Heiner

Pacific Biosciences, Menlo Park, California

The Single Molecule Real Time (SMRT®) sequencing platform provides direct sequencing data that can span several thousand bases to tens of thousands of bases in a high-throughput fashion. The capability to get very long read provides opportunities to get close-to-finished quality bacteria genome with just a single long insert (~10kb or longer) SMRTbell™ library.   Both the necessary read lengths and the necessary accuracies for successful assemblies are accomplished by new algorithmic approaches to construct accurate consensus from the reads where the dominated error modes are insertions and deletions. We demonstrate how the repeats are resolved and show the results of such process applied to assemble a bacteria genome and a previously hard-to-assemble *Plasmodium falciparum* genome.

The detection of riboswitches in stressful environments using genome-wide methods

Alexander Churkin,[1] Eviatar Nevo,[2] and Danny Barash[1]* (dbarash@cs.bgu.ac.il)

[1]Department of Computer Science, Ben-Gurion University, Beer Sheva, Israel; [2]Institute of Evolution, University of Haifa, Haifa, Israel

Riboswitches are RNA genetic control elements that were originally discovered in bacteria and provide a unique mechanism of gene regulation. They work without the participation of proteins and are believed to represent ancient regulatory systems in the evolutionary timescale. The very few eukaryotic riboswitches found so far were detected using bioinformatics and with large-scale sequencing projects underway, a new opportunity arises to perform genome-wide scans for their discovery. At JGI, peculiar organisms from stresssful environments are currently being sequenced, such as the cyanobacterium *Nostoc linckia* from "Evolution Canyon." The detection of riboswitches in *Nostoc linckia* may provide an interesting testbed to examine whether environmental stress may enhance riboswitch signals by comparing with the detection of riboswitches in cyanobacteria found in standard environments. Such a comparative study may then be used to explore the effect of environmental stress on riboswitches in higher organisms. Using specialized bioinformatic methods that have been developed for this purpose and are presented here, genome-wide scans are being performed for riboswitch detection and will also be applied to newly sequenced data.

Metasecretome phage display — a new approach for harvesting surface and secreted proteins from microbial communities

Milica Ciric* (Milica.Ciric@agresearch.co.nz),[1,2] Dragana Gagic,[1] Christina D. Moon,[1] Graeme T. Attwood,[1] Eric Altermann,[1] Sinead C. Leahy,[1] Chris J. Creevey,[3] and Jasna Rakonjac[2]

[1]AgResearch Limited, Palmerston North, New Zealand; [2]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand; [3]Teagasc, Dunsany, Ireland.

The rumen is the fermentative forestomach of ruminant animals, and one of the most studied complex microbial ecosystems in terms of targeted functional and sequence-based metagenomics. Fibrolytic activity, the most prominent feature of the rumen microbiome, depends on efficient adherence of rumen microbes to feed material, which is mediated via surface and secreted proteins - collectively referred to as the metasecretome. Consequently, the metasecretome represents a valuable repository of lignocellulolytic bioactivities and its mining will provide new information not only for improving fibre digestibility and feed efficiency in ruminant animals, but also for improving processes in cellulosic biofuel production. In order to explore the diversity of mechanisms employed by the rumen microbiome for the breakdown of plant fibre, we developed a new metasecretome phage display technology. We used this approach to specifically enrich for the rumen microbial metasecretome starting from DNA sample isolated from the plant-adherent microbial fraction of dairy cow rumen contents. This technology can be applied for harvesting and functionally investigating surface and secreted proteins from the metasecretome of any environmental sample. A pilot metasecretome phage display library was constructed to optimize the method and assess the diversity of captured metasecretome inserts by Sanger sequencing. After confirming that metasecretome selection protocol is efficient for the selection of a wide spectrum of secretome ORFs, both in terms of the diversity of membrane targeting signals present and from a taxonomically broad range of hosts, we applied this protocol to a larger shotgun library (primary size ~5 x 10^6 clones). Pyrosequencing of recombinant phagemid ssDNA isolated from virions obtained after selection was performed to fully assess the functional and taxonomic diversity of the metasecretome. The assembled and unassembled rumen microbial metasecretome datasets were automatically annotated using the JGI IMG/M pipeline. Main phylogenetic assignments, based on the distribution of best BLAST hits of predicted protein-coding genes in the assembled dataset, were to Bacteroidetes (42%) and Firmicutes (17%), while 38% remained unassigned. This finding is in agreement with predominant phyla found in the rumen. The relative abundances of Pfam groups were compared between the datasets derived from the metasecretome and metagenome. The metasecretome was enriched for genes predicted to be involved in carbohydrate transport and metabolism (19.4% compared to 10.6% in metagenomic dataset) and cell wall/membrane/envelope biogenesis. In contrast, intracellular proteins, especially within the information storage and processing functional group, were relatively less abundant. The diversity of carbohydrate-active enzyme (CAZy) families captured by metasecretome selection was estimated by dbCAN analysis of the metasecretome sequence dataset. This revealed a wide assortment of cellulases, hemicellulases, debranching enzymes and carbohydrate esterases. Notably, the dataset was enriched for the cohesin and dockerin domains, which are associated with cellulosomes – large cell-surface bound cellulose-degrading enzyme complexes that are thought to be rare in the rumen. Overall, the demonstrated enrichment for putative proteins involved in carbohydrate transport and metabolism is consistent with the involvement of the rumen metasecretome in carbohydrate metabolism coordination.

Investigating *Arabidopsis thaliana* root endophyte communities by single cell genomics

Scott Clingenpeel* (srclingenpeel@lbl.gov),[1] Derek Lundberg,[2] Tanja Woyke,[1] Susannah Tringe,[1] and Jeff Dangl[2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Land plants grow, in close association with microbial communities both on their surfaces and inside the plant (endophytes). The plant's interactions with its microbial community span the range from pathogenic to commensal or mutualistic. Colonization of the endophyte compartment occurs in the presence of a sophisticated plant immune system, suggesting finely-tuned discrimination of pathogens from mutualists and commensals. Despite the importance of the microbiome to the plant, relatively little is known about the specific interactions between plants and microbes, especially in the endophyte compartment. The vast majority of microbes have not been grown in the lab, and thus one of the few ways of studying them is by examining their DNA. Although metagenomics is a powerful tool for examining microbial communities, its application to endophytic communities is technically difficult due to the presence of large amounts of host DNA in the sample. Furthermore, metagenomics can give a picture of the genetic capabilities of the entire community, but it has rarely been able to connect specific functions to particular microbes. One method to address these difficulties is single cell genomics where a single microbial cell is isolated from a sample, lysed, and its genome amplified by multiple displacement amplification (MDA) to produce enough DNA for genome sequencing. We have applied this technology to study the endophytic microbes in *Arabidopsis thaliana* roots. Extensive 16S gene profiling of the microbial communities in the roots of multiple inbred *A. thaliana* strains has identified ~170 OTUs as being significantly enriched in all the root endophyte samples compared to their presence in bulk soil. Approximately 14,000 single microbial cells were isolated from these samples and ~500 of these were identified as being members of the enriched OTUs. The genomes of ~200 of these single cells, representing 48 of the target OTUs, are having their genomes sequenced. These genomes will be compared to those from close relatives that are not plant-associated to identify genes that are likely involved in the plant-microbe interaction.

Mining the *Agave* microbiome for adaptations to arid environments

Devin Coleman-Derr* (DAColeman-Derr@lbl.gov),**[1]** Stephen Gross,**[1]**Tanja Wojke,**[1]** Gretchen North,**[2]** Laila Partida-Martinez,**[3]** Kristen DeAngelis,**[4]** Scott Clingenpeel,**[1]** Susannah Tringe,**[1]** and Axel Visel**[1]**

[1] DOE Joint Genome Institute, Walnut Creek, California; **[2]**Department of Biology, Occidental College, Los Angeles, California; **[3]**Laboratory of Microbial Interactions, Cinvestav, Irupuato, Mexico; **[4]**Department of Microbiology, University of Massachusets, Amherst, Massachusetts

A major challenge facing the biofuels industry is the identification of high-yield plant feedstocks that can be cultivated with minimal resource inputs without competing for land and water supplies with existing food crops. Recent research has demonstrated that the *Agave* plant, cultivated in Mexico and Southwestern United States for the production of fiber and alcohol, meets these criteria[1]. Agaves grow on non-arable rocky soils in regions characterized by prolonged drought and extreme temperatures, due in part to physiological adaptions that prevent excess water-loss in arid environments[2]. Plant-microbial symbioses can play a role in helping plants adapt to heat and drought stress, increasing the accessibility of soil nutrients, or compete with plant pathogens[3]. Whether agaves have similar beneficial microbe interactions in their native environment is unknown. We aim to provide a comprehensive characterization of the *Agave* microbiome, with the goal of identifying specific community members that may contribute to *Agave* biotic and abiotic stress tolerance. We are investigating the microbial communities of both wild and cultivated *Agave* species in California and in Mexico. For each specimen, microbial samples were collected from the phyllosphere (leaf surface), rhizosphere (root surface) and leaf and root interiors (endospheres). A 16S iTag survey of these samples will be used to select specific microbiome compartments for further characterization, including shotgun metagenomics and single-cell genomics. Our project will expand our understanding of microbial diversity in desert soils, catalog and characterize the microbial factors that contribute to *Agave*'s successful adaptation to the extreme environments of its endemic range. Ultimately, we aim to enable microbiome manipulation aimed at improving the suitability of *Agave* for use in the rapidly growing biofuels industry.

[1] Davis, S., et. al. The global potential for Agave as a biofuel feedstock. *GCB BioEnergy*, 68-78, (2011).

[2] Nunez, H. et. al. Agave for tequila and biofuels: an economic assessment and potential opportunites. *GCB Bioenergy*, 43-47, (2011).

[3] Rodriguez, R. et. al. Stress Tolerance in plants via habitat-adapted symbiosis. *ISMEJ*, 404-416, (2008).

Efficient assembly of Grand Challenge metagenome data sets on a hybrid-core architecture

Alex Copeland* (accopeland@lbl.gov),[1] George Vacek,[2] Kirby Collins,[2] Susannah Tringe,[1] and Janet Jansson[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Convey Computer Corporation, Richardson, Texas

Assembly of metagenome data provides substantial benefit to downstream analysis by reducing the size and improving the accuracy of the input data, and providing contigs which are longer substrates for annotation and other analysis. To study communities with high microbial diversity containing thousands of unique genomes, metagenomes must be sequenced to sufficient depth, resulting in very large data sets. As predicted in 2009, once a sufficiently large metagenome sequence data set has been produced, it is very difficult to impossible to assemble due to substantial computational resource demands and, in particular, the very large memory requirements. Filtering the raw reads to reduce the total data set to a tractable size can lead to reduced assembly quality and is a time-consuming step in itself.

In this work, we describe using Convey's Graph Constructor (CGC) on a Convey hybrid-core (HC) server to reduce the memory and run time requirements of Velvet's assembly workflow without data reduction by a separate read filtering step. CGC accelerates construction and manipulation of de Bruijn graphs commonly used in short-read genome assembly applications and it consistently achieves memory and runtime performance gains over large shared-memory servers. Building on previous work, we present assembly results from several data sets, including the Great Prairie Grand Challenge Soil Metagenome data sets, ranging from hundreds of gigabases to over 1 terabase. The largest datasets are roughly 10 times larger than previously possible to assemble using JGI's production pipeline and completed in reasonable times using modest resources.

The DOE Systems Biology Knowledgebase: Microbial Science Domain

Paramvir S. Dehal* (psdehal@lbl.gov),[1] Christopher S. Henry (chenry@mcs.anl.gov),[2] Aaron Best,[2] Ben Bowen,[1] Steven Brenner,[1] Chris Bun,[2] Steven Chan,[1] John-Marc Chandonia,[1] Neal Conrad,[2] Matt DeJongh,[2] Paul Frybarger,[2] Keith Keller,[1] Pavel S. Novichkov,[1] Ross Overbeek,[2] Fangfang Xia,[2] Adam P. Arkin,[1] Robert Cottingham,[3] Sergei Maslov,[4] and Rick Stevens[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Argonne National Laboratory, Argonne, Illinois; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Brookhaven National Laboratory, Upton, New York

KBase is a software and data environment designed to enable researchers to collaboratively generate, test and share new hypotheses about gene and protein functions, perform large-scale analyses on our scalable computing infrastructure, and model interactions in microbes, plants, and their communities. KBase provides an open, extensible framework for secure sharing of data, tools, and scientific conclusions in the fields of predictive and systems biology. The microbes component of the KBase project aims to unify existing 'omics datasets and modeling toolsets within a single integrated framework that will enable users to move seamlessly from the genome annotation process through to a reconciled metabolic and regulatory model that is linked to all existing experimental data for a particular organism. The results are hypotheses for such things as gene-function matching and the use of comparative functional genomics to perform higher quality annotations. KBase will embody tools for applying these models and datasets to drive the advancement of biological understanding and microbial engineering. In order to drive the development of the microbes area and enable new science, we will focus on accomplishing prototype science workflows rather than general tasks. We have developed KBase workflows for: (1) genome annotation and metabolic reconstruction, (2) regulon reconstruction, (3) metabolic and regulatory model reconstruction, and (4) reconciliation with experimental phenotype and expression data.

Genomic analysis of *Zymomonas mobilis* subsp. *mobilis* ATCC 29191: Comparative, structural and functional insights

Andreas Desiniotis,[1] Vassili N. Kouvelis,[1] Karen Davenport,[2] David Bruce,[3] Chris Detter,[2] Roxanne Tapia,[2] Cliff Han,[2] Miriam L. Land,[4] Loren Hauser,[4] Yun-Juan Chang,[4] Chrongle Pan,[4] Lynne A. Goodwin,[2] Tanja Woyke,[2] Nikos C. Kyrpides,[3] Milton A. Typas,[1] and Katherine M. Pappas*[1] (kmpappas@biol.uoa.gr)

[1]Department of Genetics & Biotechnology, Faculty of Biology, University of Athens, Panepistimiopolis, Athens, Greece; [2]DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Oak Ridge National Laboratory, Bioscience Division, Oak Ridge, Tennessee

Zymomonas mobilis is an α-proteobacterium with a great potential for biofuel production as it ferments sugars to ethanol, to almost perfect yields. Different strains of *Z. mobilis* are being sequenced at the US Department of Energy Joint Genome Institute in collaboration with the University of Athens, Greece (CSP_788284; CSP_52). The most recently sequenced *Z. mobilis* strain is the phenotypic centrotype of the subspecies *mobilis*, strain ATCC 29191 (Desiniotis *et al., J. Bacteriol*. 194; 5966-7). ATCC 29191, originally isolated from palm wine fermentations in Congo, Africa, has been extensively studied in the past as a model for respiration and energy conversion in the genus, and among other *Z. mobilis* strains is also reported to be superior in levan (polyfructan) production. The genome of ATCC 29191 comprises a

circular chromosome of 1,961,307-bp size, and three plasmids of 18,350-bp, 14,947-bp and 13,742-bp size, respectively. The entire genome has 1,765 protein-coding genes, 3 rRNA clusters and 51 tRNAs. Genomic comparisons between ATCC 29191 and ATCC 31821 (ZM4) used as reference, revealed that the former is 95,057 bp smaller, bears over 40 strain-specific genes (compared to 120 for ZM4), and displays genomic rearrangements, many of which are due to an impressive number of 24 chromosome-borne and 4 plasmid-borne insertion elements (ISs). Almost all ISs belong to the IS*4* family and are divided into two subfamilies, depending on the presence or absence of specific domains in transposase ORFs. ATCC 29191 genes that exhibited strain specificity compared to ZM4, as well as all hypothetical ATCC 29191 genes, were curated and sought in the genomes of hitherto sequenced and deposited *Z. mobilis* strains. Among well designated strain-specific genes –chromosomal or plasmid– included were genes coding for tellurium resistance, and also metabolic, regulatory, and DNA biosynthesis/relocation genes. Due to the fact that ATCC 29191 is a robust *Z. mobilis* levan producer, an effort was made to determine all structural and regulatory genes that contribute to sucrose uptake and hydrolysis, and fructose polymerization.

DNA synthesis and assembly pipeline at the Joint Genome Institute

Sam Deutsch (sdeutsch@lbl.gov),[1] Sarah Richardson,[1] Angela Tarver,[1] Matthew Hamilton,[1] David Robinson,[1] Sangeeta Nath,[1] Matthew Bendall,[1] Miranda Harmon-Smith,[1] Lisa Simirenko,[1] Nathan J. Hillson,[1,2] and Jan-Fang Cheng[1] (jfcheng@lbl.gov)

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Joint BioEnergy Institute, Emeryville, California

Next generation sequencing has generated a wealth of DNA sequence information. Towards bridging the gap between digital sequence information and functional biological understanding, we have developed a production DNA synthesis and assembly pipeline to construct biological parts, pathways and even artificial chromosomes *de novo*. The pipeline begins with a suite of informatics tools that assist the design of genetic components (codon optimization, partitioning, oligo design) their arrangement into larger constructs, and potentially strategies for generating combinatorial libraries. We have transferred most steps of the pipeline to robotic workstations including oligo liquid-handling oligo assembly steps, transformant plating and colony picking. We are continuously improving our procedures to achieve higher throughput and lower cost. For example, we have recently employed the Labcyte Echo liquid dispenser to reduce assembly reaction volumes 10-fold without impacting product quality. Assembled DNA constructs are sequence validated with the PacBio platform, whose long sequencing read-lengths enable the resolution of multi-kilobase constructs within a combinatorial library. Sequence validated clones are shipped to users as DNA and/or glycerol stocks or used for functional characterization in-house. In FY12, we produced 1.5 Mb of synthetic DNA, and we are on track to produce 3.5 Mb in FY13. We anticipate that this synthesis capacity will continue to expand as new technologies are developed and implemented.

Longer read length, higher throughput SMRT® sequencing

John Eid* (jeid@pacificbiosciences.com), Paul Peluso, David Rank, Satwik Kamtekar, Erik Miller, Walter Lee, Phil Jiao, Colleen Cutcliffe, Paul Lundquist, Annette Grot, Michael Gandy, Jeremiah Hanes, Keith Bjornson, Lubomir Sebo, Louis Brogley, Regina Lam, Gene Shen, Honey Osuna, Anushweta Asthana, Arunashree Bhamidipati, Emilia Mollova, Alicia Yang, John Lyle, Joan Wilson, Kevin Travers, and Michael Phillips

Pacific Biosciences, Inc., Menlo Park, CA

Pacific Biosciences has developed an even longer sequencing chemistry that generates mean read lengths of >4,000 bases, with 5% of reads achieving lengths of >12,000 bases. These read length advances put the PacBio sequencing technology at the limits of current single-molecule, long-template preparation capability and so methods to size select have been developed that enhance these read length gains. Data on an even more advanced, next-generation chemistry, using the PacBio photo-protected approach has also progressed. This new approach, which is still in development, in conjunction with the advances in template size-selection, yields mean read lengths of 8,000 bases, with 5% of reads producing greater than 18,000 bases of contiguous sequence. In addition, the number of ZMWs that can be simultaneously interrogated has been doubled from 75,000 to 150,000. The combination of these improvements will result in a daily instrument throughput of 2-3 Gb. This throughput level, combined with the increasingly long read length capabilities of the system will enable unique sequencing applications on mammalian size genomes.

High throughput miniaturized PCR using the Echo® liquid handler

Celeste Glazer* (cglazer@labcyte.com), Joe Barco, Brent Eaton, and Sammy Datwani

Labcyte Inc., Sunnyvale, California

Quantitative PCR (qPCR) has become widely prevalent across many disciplines throughout drug discovery. Advances in qPCR detection technology to include 384- and 1536-well plate formats have enabled researchers to increase throughput while decreasing reagent costs. To ensure high data quality, the liquid handling employed in such low-volume reactions must be precise and accurate. Tipless, touchless acoustic droplet ejection with the Echo® liquid handler eliminates the cost of disposable tips and wash cycles and improves workflow by simplifying assay setup. This study utilized new transfer methodologies to greatly reduce transfer times. Precision for the resulting quantification curves across 384- and 1536-well plates was excellent with standard deviations less than 0.25 and CVs less than 2.0%. Tests with positive and negative controls dispensed into alternating wells revealed zero cross-contamination. The results confirm the advantages of using the Echo liquid handler for preparing high-throughput miniaturized qPCR in both 384- and 1536-well formats.

Application of parallel promoter stacking to the study of effector-triggered immunity in plants

Tania L. Gonzalez[1] and Ming C. Hammond* (mingch@berkeley.edu)[1,2]*

[1]Department of Molecular & Cell Biology, [2]Department of Chemistry, University of California at Berkeley, Berkeley, California

Here we describe a new approach to transgene design called "parallel promoter stacking" that enables two or more promoters to be combined in a plug-and-play fashion to regulate any gene of interest. This design enables multi-parameter regulation without the need for specific transcription factors or promoter engineering. We showcase the ability of promoter stacking to overcome the common problem of leaky expression from chemically inducible promoters by enforcing tight control of the expression of a bacterial effector protein. Our strategy enables non-leaky, inducible expression of single effector proteins in transgenic plants, which will allow for the elucidation of effector-triggered immunity signaling pathways.

Genome diversity in *Brachypodium distachyon*: Deep sequencing of highly diverse natural accessions

Sean P. Gordon* (Sean.Gordon@ars.usda.gov),[1] Henry Priest,[2] David L. Des Marais,[3] Wendy Schackwitz,[4] Melania Figueroa,[5] Joel Martin,[4] Jennifer N. Bragg,[1] Ludmila Tyler,[6] Cheng-Ruei Lee,[7] Doug Bryant,[2] Wenqin Wang,[8] Joachim Messing,[8] Kerrie Barry,[4] David Garvin,[9] Hikmet Budak,[10] Metin Tuna,[11] Thomas Mitchell-Olds,[7] William F. Pfender,[5] Thomas Jeunger,[3] Todd C. Mockler,[2] and John P. Vogel[1]

[1]U.S. Department of Agriculture-Agricultural Research Service-Western Regional Research Center, Albany, California; [2]Donald Danforth Plant Science Center, Saint Louis, Missouri; [3]University of Texas at Austin, Austin, Texas; [4]DOE Joint Genome Institute, Walnut Creek, California; [5]U.S. Department of Agriculture-Agricultural Research Service, Corvallis, Oregon; [6]University of Massachusetts, Amherst, Massachusetts; [7]Duke University, Durham, North Carolina; [8]Waksman Institute, Rutgers University, Piscataway, New Jersey; [9]U.S. Department of Agriculture-Agricultural Research Service, Plant Science Research Unit, St. Paul, Minnesota; [10]Sabanci University, Istanbul, Turkey; [11]Namik Kemal University, Tekirdag, Turkey

Natural variation is a powerful resource for studying the genetic basis of biological traits. Brachypodium distachyon (Brachypodium) is a model grass with a small genome and a large collection of diverse, inbred, diploid lines. As a step towards understanding the genetic basis for this variation, we sequenced the reference accession, Bd21, and 6 divergent strains to 34- to 58-fold coverage with paired-end Illumina reads. We identify a total of 5,216,105 unique SNPs among divergent accessions and generate a subset of 2,485,097 high-confidence non-redundant SNPs for mapping purposes, containing 96.6% of SNPs previously used to produce a genetic linkage map. We identified over 2,000 potential errors in the reference genome, and correct 170 gaps. We further identify > 1,000,000 small and large indels that account for a non-redundant 11.3Mb of sequence inserted or deleted in one or more accessions. We have generated assemblies and gene annotations for the six divergent lines. In addition, we produced a *de novo* transcriptome for the most divergent line revealing >2,400 transcripts not in the reference annotation. We use RNA-Seq from the 6 divergent lines and the Bd21 reference to functionally validate genomic variant predictions. In addition, we show large-scale differences in expression responses to water deficit. Download and visualization of our data is available at http://www.brachypodium.org.

Genome-level diversity of nitrogen-fixing symbionts of chickpea: Insights into the effect of domestication on symbiotic nitrogen fixation

Alex Greenspan* (greenspan@ucdavis.edu),[1] Donghyun Kim,[2] Aamir Khan,[2] Rajeev Varshney,[2] and Douglas Cook[1]

[1]Department of Plant Pathology, University of California, Davis,Davis, California; [2]International Crop Research Institute for the Semi-Arid Tropics, Hyderabad, India.

Legume crops are significant for their ability to host a diverse group of nitrogen-fixing bacterial symbionts, collectively called rhizobia. Reduced nitrogen is the growth-limiting nutrient in most agricultural systems. Modern agricultural practices rely on industrially produced, synthetic nitrogenous fertilizers to alleviate this limitation, but the production and application of these fertilizers contributes substantially to greenhouse gas emissions and environmental degradation. Legume crops provide an important alternative. The legume-rhizobium symbiosis contributes as much as 40 million tons of fixed-nitrogen to agricultural soils yearly (approximately half of global consumption of synthetic nitrogenous fertilizer). However, there is evidence to suggest that crop legumes are less efficient at fixing nitrogen than their wild relatives. The mechanisms underlying these differences have not been addressed. This study attempts to examine differences in nitrogen fixation between wild and domesticated chickpea, by using whole-genome sequencing to characterize the population of chickpea's rhizobial symbionts (*Mesorhizobium spp.*). The legume-rhizobium symbiosis occurs in plant-derived root organs called root nodules, and is generally highly specific with one species of legume pairing with one or a few lineages of rhizobia. Our hypothesis was that crop breeding led to altered symbiotic specificity in domesticated chickpea relative to its wild progenitor. To evaluate this, we extracted DNA from root nodules collected from wild (*Cicer reticulatum* and *Cicer echinosperum*) and domesticated (*Cicer arietinum*) chickpea, grown in agricultural fields in. We sequenced total DNA from 95 nodules using Illumina's Hi-Seq 2000. Genome assemblies from nodule DNA allow us to infer population genetic relationships of *Mesorhizobia* inhabiting nodules using nucleotide polymorphisms in core-genome sequences and gene presence/absence variation in the *Mesorhizobium* accessory genome. In addition, sequencing nodule DNA also gives us unprecedented perspective on the prevalence of mixed infections by multiple *Mesorhizobium* strains that would not be detected in culture, as well as other bacteria occurring in high abundance in mature nodules. After filtering out DNA reads that aligned to the *C. arietinum* genome, we were able to assemble *Mesorhizobium* genomes from 42 out of 95 nodules, which we can use to compare symbiont populations across chickpea species. Also, using approaches developed for metagenomic analysis of bacterial communities, we found that in 13 of 95 nodules, the most prevalent bacterium belonged to the genus *Pseudomonas.* Future work will include developing a scalable PCR assay from the accessory components of our assembled draft genomes to rapidly characterize *Mesorhizobium* strains in nodules. This will allow us to map genetic differences in wild and domesticated chickpea that underlie differences in symbiotic specificity. We will also characterize the *Mesorhizobium* population occurring in *Cicer reticulatum* root nodules in the plant's native range in Southern Turkey.

Fungal Genomics for Energy and Environment

Igor Grigoriev* (ivgrigoriev@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

Genomes of fungi relevant to energy and environment are in focus of the Fungal Genomic Program at the US Department of Energy Joint Genome Institute (JGI). One of its projects, the Genomics Encyclopedia of Fungi, targets fungi related to plant health (symbionts, pathogens, and biocontrol agents) and biorefinery processes (cellulose degradation, sugar fermentation, industrial hosts) by means of genome sequencing and analysis. New chapters of the Encyclopedia can be opened with user proposals to the JGI Community Sequencing Program (CSP). Another JGI project, the 1000 fungal genomes, explores fungal diversity on genome level at scale and is open for users to nominate new species for sequencing. Over 200 fungal genomes have been sequenced by JGI to date and released through MycoCosm (www.jgi.doe.gov/fungi), a fungal web-portal, which integrates sequence and functional data with genome analysis tools for user community. Sequence analysis supported by functional genomics leads to developing parts list for complex systems ranging from ecosystems of biofuel crops to biorefineries. Recent examples of such 'parts' suggested by comparative genomics and functional analysis in these areas are presented here.

Transcriptome analysis of drought tolerant CAM plants, *Agave deserti* and *A. tequilana*

Stephen M. Gross* (smgross@lbl.gov),[1,2] Jeffrey A. Martin,[1,2] June Simpson,[3] Zhong Wang,[1,2] and Axel Visel[1,2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]Cinvestav, Irapuato, Mexico

Agaves are succulent monocotyledonous plants native to hot and arid environments of North America. Because of their adaptations to their environment, including crassulacean acid metabolism (CAM, a water-efficient form of photosynthesis) and existing technologies for ethanol production, agaves have gained attention both as potential lignocellulosic bioenergy feedstocks and models for exploring plant responses to abiotic stress. However, the lack of comprehensive *Agave* sequence datasets limits the scope of investigations into the molecular-genetic basis of *Agave* traits. Here, we present comprehensive, high quality *de novo* transcriptome assemblies of two *Agave* species, *A. tequilana* and *A. deserti,* from short-read RNA-seq data. Our analyses support completeness and accuracy of the *de novo* transcriptome assemblies, with each species having a minimum of 35,000 protein-coding genes. Comparison of agave proteomes to those of additional plant species identifies biological functions of gene families displaying sequence divergence in agave species. Additionally, we use RNA-seq data to gain insights into biological functions along the *A. deserti* juvenile leaf proximal-distal axis. Our work presents a foundation for further investigation of agave biology and their improvement for bioenergy development.

Towards finished genome assemblies using SMRT® sequencing

Jenny Gu* (jgu@pacificbiosciences.com), Lawrence Hon, Jason Chin, David Alexander, and Aaron Klammer

Pacific Biosciences, Menlo Park, California

Plant and animal genomes are replete with repetitive sequences that pose challenges for short-read-based assemblies. By using reads greater than 5 kb, SMRT® Sequencing can span many of these repeats, which allows the joining of contigs that would otherwise be split at a repeat, thereby improving assembly continuity. Two recent algorithm advances increase the utility of SMRT Sequencing data. First, the hierarchical genome assembly (HGA) process reduces library requirements to a single 10 kb or longer SMRTbell™ library by simultaneously utilizing the information from both longer and shorter reads. The approach enables a more efficient de novo assembly by performing a pre-assembly step that generates highly accurate, consensus-based reads that can be used as input for existing genome assemblers. Second, the Quiver consensus caller is a hidden Markov model-based method that can be used to polish assemblies, resulting in per-base accuracy of >QV50. By modeling random insertions and deletions, which are more common in single molecule sequencing, the algorithm can estimate near-perfect consensus sequences. We demonstrate these methods in a number of examples.

Analysis of microbial communities associated with natural oils that seep into the Santa Barbara Channel: Linking community dynamics with biological hydrocarbon degradation

Erik R. Hawley*,[1] Thomas D. Lorenson,[2] and **Matthias Hess** (matthias.hess@tricity.wsu.edu)[1,3,4]

[1]Washington State University, School of Molecular Biosciences, Systems Biology & Applied Microbial Genomics Group, Pullman, Washington; [2]U.S. Geological Survey, Menlo Park, California; [3]Pacific Northwest National Laboratory, Chemical and Biological Process Development Group, Richland, Washington; [4]Pacific Northwest National Laboratory, Environmental Molecular Sciences Laboratory, Richland, Washington

Marine microbial consortia are known for their ability to degrade hydrocarbons. Employing the intrinsic bioremediation potential of hydrocarbon-degrading microbes has been used successfully in the remediation of oil that contaminated long stretches of shorelines and it was endorsed anew as promising remediation strategy after the Deepwater Horizon blowout. Despite the significant resources that have been spent to study the microbial response to oils spills, most of the research in this field comes from culture-based studies and relatively little is known about the dynamics and microbial processes that occur during the biological degradation of crude oil in uncontrolled and highly complex biological systems. To improve our understanding of the microbiology and the community dynamics associated with natural oils that enter the marine ecosystem, we amplified the 16S rRNA gene from the microbial community associated with five crude oils that seep into the Santa Barbara Channel offshore Southern California. Community profiles were generated using pyrosequencing and a DNA microarray approach and the obtained results suggest that organisms, for which no cultured representatives have been describes until today, dominate the communities associated with these hydrocarbons. Furthermore, analysis of the ~39,000 sequence reads generated during this project suggest that hydrocarbon-degrading communities found in the Santa Barbara Channel differ from the ones found in the Gulf of Mexico and that have been reported to have played a major role in the rapid degradation of

the oil that was polluted the marine ecosystem in 2010. In summary, the results presented here suggest that distinctive microbial communities, dominated by uncultured representatives, might be responsible for the degradation of natural hydrocarbons in the Santa Barbara Channel and the Gulf of Mexico.

Greater than 10 kb read lengths routine when sequencing with Pacific Biosciences' XL release

Cheryl Heiner* (cheiner@pacificbiosciences.com), Primo Baybayan, Susana Wang, Meredith Ashby, Yan Guo, and Jason Underwood

Pacific Biosciences, Menlo Park, California

PacBio's SMRT® Sequencing produces the longest read lengths of any sequencing technology currently available. There have been a number of recent improvements to further extend the length of PacBio® *RS* reads. With an exponential read length distribution, there are many reads greater than 10 kb, and some reads at or beyond 20 kb. These improvements include library prep methods for generating >10 kb libraries, a new XL polymerase, magnetic bead loading, stage start, new XL sequencing kits, and increasing data collection time to 120 minutes per SMRT Cell. Each of these features will be described, with data illustrating the associated gains in performance.

With these developments, we are able to obtain greatly improved and, in some cases, completed assemblies for genomes that have been considered impossible to assemble in the past, because they include repeats or low complexity regions spanning many kilobases. Long read lengths are valuable in other areas as well. In a single read, we can obtain sequence covering an entire viral segment, read through multi-kilobase amplicons with expanded repeats, and identify splice variants in long, full-length cDNA sequences. Examples of these applications will be shown.

Lignocellulose bioconversion by engineered mixed cultures

Guillermina Hernandez-Raquet* (hernandg@insa-toulouse.fr), Sandrine Païssé, Adèle Lazuka, Lucas Auer, Sophie Bozonnet, and Michael O'Donohue

Université de Toulouse, INSA, UPS, LISBP, Toulouse, France ; INRA, Ingénierie des Systèmes Biologiques et des Procédés, Toulouse, France; CNRS, UMR, Toulouse, France

The bioconversion of lignocellulosic (LC) waste for the production of energy and chemicals is a scientific and economic challenge. LC is deconstructed using physico-chemical treatments combined to enzymatic hydrolysis and fermentation by selected microbial strains. However, the use of expensive enzymatic treatments and single strains displaying limited enzymatic potential to transform complex LC-waste is source of debate. An alternative is the development of engineered mixed culture (EMC) processes to transform LC-waste into fermentation products that can be further processed into energetic or value-added products. In nature, LC is degraded by various enzyme families acting synergistically. So, using EMC can increase the metabolic diversity needed for LC transformation. Moreover, EMC display a high stability in broad range of conditions and can be used in non-sterilized conditions. The rumen and thermite gut ecosystems are particularly interesting source of inoculum to obtain hyper-hydrolytic MC, because these ecosystems produce a large panel of LC degrading enzymes. Previous studies using EMC to transform LC have been mainly focused on macro-kinetic parameters (e.g. productivity). But, little information exists about the microbial diversity present in such engineered ecosystems. Our objective is to study the functional diversity of engineered microbial communities displaying high LC deconstructing capabilities. For the engineered ecosystems, different reactors fed with wheat straw pretreated in different ways (alkaline, oxidative and ammonia pretreatment) will be compared for their macro-kinetic parameters (transformation rate, productivity). The functional diversity of these ecosystems will be analyzed by pyrosequencing of 16SrRNA. Our objective is to correlate the data obtained from functional diversity to those obtained from substrate composition, transformation rates and enzymatic activity. The engineered consortium displaying high hydrolytic potential will be selected for the metagenomic and metatranscriptomic studies.

Assessment of Nextera™ long mate-pair libraries: A rapid, low-input method for mate-pair library construction yields improved assemblies

Cindi A. Hoover* (cahoover@lbl.gov), Kevin S. Eng, Hui Sun, Jeff Froula, and Feng Chen

DOE Joint Genome Institute, Walnut Creek, California

Long mate-pair libraries are invaluable tools for genome assembly. However, traditional methods of long mate-pair library construction require large (20 µg) quantities of DNA and several days of hands-on time. Illumina's Nextera™ Long Mate-Pair (LMP) method is rapid and requires only 1 to 4 micrograms of input material. Here we present an initial assessment of the method for both gel-free and gel size-selected libraries using microbial, fungal, and plant samples. We observed uniform read coverage and high read uniqueness for Nextera™ LMP libraries. Assembly using ALLPATHS-LG generated low contig and scaffold numbers even with relatively low mate-pair coverage.

Deep community sequencing using short Illumina and multi-kb Moleculo reads provides insight into aquifer sediment microbial diversity and metabolic potential

Laura A. Hug* (laura.hug@berkeley.edu),[1] Itai Sharon,[1] Cindy J. Castelle,[1] Brian C. Thomas,[1] Kelly C. Wrighton,[1] Kenneth H. Williams,[2] Susannah G. Tringe,[3] and Jillian F. Banfield[1]

[1]Department of Earth and Planetary Science, University of Calfornia at Berkeley, Berkeley, California; [2]Geophysics Department, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]Metagenome Program, DOE Joint Genome Institute, Walnut Creek, California

Sediments are massive reservoirs of carbon compounds and host a large fraction of the microbial life on earth. Despite this, relatively little is known about sediment microorganism communities and their functions within carbon, nitrogen, and hydrogen cycling in the subsurface. We conducted deep metagenome sequencing of sediment samples from three different depths in a freshwater aquifer adjacent to the Colorado River, USA. The complete sequence dataset comprised four lanes of Illumina HiSeq and three Moleculo/Illumina libraries. Methodologies for utilizing a hybrid dataset of multi-kb Moleculo/Illumina and short HiSeq reads were developed. Phylogenetic profiling of the abundant organisms depicts a highly diverse, low-abundance community, where no single organism constituted more than 1% of the total community. Many of the bacterial and archaeal lineages sampled are highly novel, including fifteen previously genomically un-sampled phyla. The Moleculo sequencing resulted in genomic information for organisms at even lower abundance in the community compared to the Illumina assemblies, and provided a mechanism for curation of reconstructed genomes. A complete genome was reconstructed for the dominant organism in the 5m depth sample, which represents a novel bacterial phylum. The substantial evolutionary distance separating this novel bacteria phylum

from well-characterized organisms resulted in identification of significant enzymatic novelty. Draft genomes were reconstructed for three phylogenetically distant Chloroflexi species, an abundant but understudied phylum in the subsurface. The metabolic potential deduced from all four of the reconstructed genomes reveals roles for these organisms in carbon fixation and cycling, as well as mechanisms for adapting to fluctuating redox conditions near the water table. The results provide new insight into the scale of microbial diversity in the subsurface and indicate the importance of undescribed and understudied groups in sediment biogeochemical transformations.

The ecology and evolution of secondary metabolite gene collectives in a marine Actinomycete lineage

Paul R. Jensen* (pjensen@ucsd.edu), Krystle Chavarria, Natalie Millan, Anna Lechner, and Nadine Ziemert

Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

Secondary metabolites are produced by large gene collectives that are well known to be exchanged by horizontal gene transfer. The small molecule products of these pathways often possess potent biological activities that can improve fitness, yet their evolutionary histories and the extent to which acquisition events are associated with speciation or periodic selection has not been explored. We are addressing these questions in the marine actinomycete genus *Salinispora*, a rich source of biologically active secondary metabolites. As part of an ongoing JGI/CSP project, we are analyzing the genome sequences of >100 *Salinispora* strains derived from eight global collection sites including multiple representatives of different species derived from the same collection site. This research is providing new opportunities to address the question: what's more important who you are or where you live, in the context of secondary metabolism and genome evolution. We have documented extraordinary rates of pathway sampling, which provide new clues for bioprospecting strategies. We have also documented species-specific pathway distributions suggesting that secondary metabolites represent ecotype defining traits that appear tied to period selection events. Pathway evolutionary histories are revealing the processes that generate new small molecule diversity and the extent to which these pathways are exchanged among closely related species.

Nitrogen metabolism in plants and green algae: A case study using *Micromonas*

Valeria Jimenez* (vjimenez@mbari.org),[1] Chia-Lin Wei,[2] Chee-Hong Wong,[2] Chew Yee Ngan,[2] and Alexandra Z. Worden[1]

[1]Monterey Bay Aquarium Research Institute, Moss Landing, California; [2]Sequencing Technology Group, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California

Micromonas is a wide spread marine alga that belongs to the Plantae. Prasinophytes like *Micromonas* are distinct from chlorophyte algae (e.g. *Chlamydomonas*) and land plants and taken together these three lineages provide insights to the Plant ancestor. Two species of *Micromonas* represented by isolates CCMP1545 and RCC299 have been completely sequenced. We investigated genes involved in nitrogen metabolism and differences between the two species, as well as other algae and plants. These include multiple different types of ammonium (AMT) transporters, previously inferred to represent *Micromonas* optimization for uptake from the environment. Because specific cellular roles for these genes are still unclear, we performed growth experiments in nutrient replete conditions over the 14:10 hr light dark (diel) cycle as well as nitrogen limitation experiments. In addition to growth rate and pigment measurements, stranded RNAseq was performed to analyze gene expression. Similar to results for many other *Micromonas* genes, differential expression was observed for most nitrogen related genes over the diel cycle (for both strains). Nitrate (NRT and NAR2), nitrite (NAR1) and AMT transporters were up-regulated at the end of the night period and during the first part of the light period. One AMT type is highly up-regulated relative to others whereas a preliminary analysis indicates that a different AMT type is highly up-regulated under nitrogen starvation. Likewise our analysis revealed variations in NRT, NAR2 and NAR1, glutamine and glutamate synthases, asparagines synthase, aspartate amino transferase and molybdenum, molybdopterin and molybdate synthesis (cofactors), and other genes involved in nitrogen metabolism associated with the cell cycles as opposed to nitrogen stress. Our study advances knowledge of the ancestral component of nitrogen metabolism genes in the green lineage. Moreover it enhances knowledge of nitrogen metabolism in marine algae, particularly with respect to factors that expressed in relation to acquisition of nitrogen from the environment, and nitrogen stress, versus intracellular trafficking and the cell cycle.

SIGMA: A Bayesian model based clustering approach for reconstructing individual genomes from shotgun sequencing of microbial communities

M. Senthil Kumar,[1,2] **Denis Bertrand**,[1] Song Gao,[3] and Niranjan Nagarajan* (nagarajann@gis.a-star.edu.sg)[1]

[1]Computational and Systems Biology, Genome Institute of Singapore, Singapore; [2]Graduate Program in Computational Biology and Bioinformatics, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore

For complex microbial communities, the analysis of information rich, whole-community shotgun sequencing datasets is often limited by the fragmentary nature of the assembly. While recent work has shown that near complete genomes can be assembled from metagenomic data, a robust and fully-automated approach to consistently do so has yet to be established. In this work, we show that the problem of metagenomic assembly can be accurately and automatically reduced to the single genome assembly problem by systematically exploiting genome coverage and assembly information. To do this, we introduce a model-based clustering approach called SIGMA that finds an optimal solution based on the Bayes Information Criterion (BIC). We combined this method with an optimal single-genome scaffolder (Opera) to show that near-complete genomes can be automatically re-constructed even from shotgun sequencing of complex microbial communities. Comparisons on *in silico* and real datasets confirmed that our approach (OperaMS) consistently outperforms state-of-the-art single genome (Velvet, SOAPdenovo) and metagenomic (MetaVelvet, Bambus2) assembly tools. Availability: https://sourceforge.net/projects/operams/

Genome assembly improvement with Pacific Biosciences RS long-read sequencing technology

K. LaButti* (klabutti@lbl.gov), A. Copeland, A. Clum, and H. Sun

DOE Joint Genome InstituteWalnut Creek, California

Advances in sequencing technology and assembly methods have allowed genomes to be sequenced and assembled at vastly decreased costs. However, the assembled data is frequently fragmented with many gaps. In the past assembly improvement was performed using time-consuming and costly Sanger-based manual finishing processes. Here we present a comparison of two methods, AllPaths-LG and PBJelly, which employ long read data from the Pacific Biosciences RS platform to improve Illumina draft assemblies in an automated fashion.

Delineating molecular interaction mechanisms in an *in vitro* microbial-plant community

Peter E. Larsen* (plarsen@anl.gov),[1] Shalaka Desai,[1] Sarah Zerbs,[1] Avinash Sreedasyam,[2] Geetika Trivedi,[2] Leland J. Cseke,[2] and Frank R. Collart[1]

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

In a forest ecosystem, the aboveground environmental phenotype (e.g. biomass, carbon allocation, photosynthetic rates) is dependent upon the below ground community interactions. The subsurface soil community comprised of fungi, bacteria, and plant roots can be modeled as compartmentalized metabolisms connected to one another and their environment via nutrient exchange and sensor networks. The sensory networks comprise a complex set of interactions between membrane receptors, signal cascade proteins, transcription factors, and transcription factor biding DNA motifs which drive gene expression patterns in response to specific combinations of extracellular signals. We have designed specific biological experiments designed to isolate the mechanisms and signaling components of interactions for all pair-wise combinations of *in vitro* soil community. Our *in vitro* laboratory soil community system is comprised of the tree *Populus tremuloides* (aspen), the ectomycorrhizal symbiotic fungus *Laccaria bicolor* (Laccaria), and four strains of the bacteria *Pseudomonas fluorescens* (Pfl0-1, Pf-5, SBW25, and WH6). Possible interactions are assessed for various permutation of aspen root with Laccaria exudate, *P. fluorescens* with root exudate, and Laccaria with *P. fluorescens* exudate. We present an example of our analysis pipeline for modeling sensor complexes using the aspen root/Laccaria exudate experimental results. As these systems are comprised of many elements, predicting sensor mechanisms requires approaches that link genomic, transcriptomic, proteomic, and metabolomic modeling. The model is derived using *in vitro* experiments of Laccaria and aspen interaction with over 200 previously published aspen root transcriptomic experimental data sets, more than 120K experimentally observed protein-protein interactions, 159 experimentally validated plant transcription factor binding motifs, and the sequenced and annotated genomes of aspen and Laccaria. Fifteen transcription factors and 13 trans-membrane receptors, in 4 protein interaction sensor complexes, are predicted to regulate the expression of 1184 genes in response to specific metabolites synthesized by Laccaria. These metabolites include phenylpropanoids, salicylate, and, jasmonic acid. Eight specific transcription factor DNA binding motifs in the aspen genome are identified as important to pre-mycorrhizal interaction and link sensor mechanisms to regulatory mechanisms. Application of this analysis pipeline will provide similar models of molecular mechanisms in fungus and *P. fluorescens* for

validation. Results of these pair-wise interaction systems will greatly reduce the search-space necessary to model interaction mechanisms in the complete, three-member *in vitro* community model.

Improving biofuel feedstocks by modifying xylan biosynthesis

Jane Lau* (Jlau@lbl.gov),[1] Pia Damm Petersen,[1,2] Jin Sun Kim,[1] Fan Yang,[1] Yves Verhertbruggen,[1] Berit Ebert,[1] Patanjali Varanasi,[1] Anongpat Suttangkakul,[1] Manfred Auer,[1] Dominique Loque,[1] and Henrik Vibe Scheller[1,3]

[1]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Department of Plant Biology and Biotechnology, University of Copenhagen, Frederiksberg, Denmark; [3]Department of Plant & Microbial Biology, University of California at Berkeley, Berkeley, California

Plant biomass for bioenergy purposes is composed largely of secondary cell walls, about a third of which is hemicellulose. In angiosperms, hemicellulose is mainly composed of xylans, which are polymers pentoses that are less desirable than hexoses for fermentation. Unfortunately, these polymers cannot be easily removed without impacting cell wall strength. Thus, plants with low xylan content have collapsed vessels which compromises their water and nutrient transport function. In order to improve plant biomass quality with optimized hexose/pentose ratio, we developed methods to spatially and temporally fine-tune the deposition of xylan by using specific transcription factors. We have generated transgenic Arabidopsis plants where xylan is synthesized normally in vessels but not in interfascicular fibers. With this approach, we generated healthy plants with reduced xylan that resulted in improved properties for saccharification and production of more easily fermentable sugars. The engineering of reduced xylan can be combined with fine-tuning of lignin and increased deposition of C6 sugars in the fibers.

Modulation of root microbiome community assembly by the plant immune response

Sarah L. Lebeis* (lebeis@live.unc.edu),[1] **Sur Herrera Paredes*** (sur00mx@gmail.com),[1] Derek S. Lundberg,[1] Natalie W. Breakfield,[1] Scott Yourstone,[1] Jase Gehring,[1,2] Stephanie Malfatti,[3] Scott Clingenpeel,[3] Tijana Glavina del Rio,[3] Philip Hugenholtz,[3,4] Susannah Green Tringe,[3] and Jeffery L. Dangl[1]

[1]Department of Biology, University of North Carolina, Chapel Hill, North Carolina; [2]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences & Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia

Multicellular eukaryotes have evolved in a microbial world with many harnessing benefits from their surroundings via assembly of defined microbial communities. While soils are an extremely diverse microbial environment, a narrower subset of microbial taxa lives within root tissue, suggesting finely-tuned discrimination by the plant host. Here, we demonstrate that the plant immune system modulates root microbiome composition. We used massive parallel sequencing of the ribosomal 16S gene to profile the root-associated microbial communities of Arabidopsis mutants with either constitutive (Hyper-) or impaired (Hypo-) immune phytohormone production and/or signaling. Although overall root microbiome composition is remarkably consistent, both Hyper- and Hypo-immune plants display significantly altered Actinobacteria, Firmicutes, and Proteobacteria abundances compared to their

isogenic wild-type controls. Our analysis of the data with a Zero-Inflated model identified specific taxa associated with each plant genotype. We have representative bacterial isolates of these differential taxa to use in reconstructed microcosms experiments to confirm these associations. We are also in the process of confirming our results with a different sequencing platform (Illumina MiSeq), and with a PCR-independent method (CARD-FISH). Together, we demonstrate that host immune response coordinates root microbiome composition, potentially tailoring host health and maximizing plant productivity across complex microbial environments.

Metagenomic and metatranscriptomic sequencing of a hypersaline mat diel

Jackson Z. Lee* (Jackson.z.lee@nasa.gov),[1,3] Angela M. Detweiler,[1,3] Mike D. Kubo,[1,4] R. Craig Everroad,[1] Tori M. Hoehler,[1] Peter K. Weber,[2] Jennifer Pett-Ridge,[2] Leslie E. Prufert-Bebout,[1] and Brad M. Bebout[1]

[1]NASA Ames Exobiology Branch, Moffett Field, California; [2]Lawrence Livermore National Laboratory, Livermore, California; [3]Bay Area Environmental Research Institute, Sonoma, California; [4]The SETI Institute, Mountain View, California

The bacterial communities of hypersaline microbial mats are an ideal complex model microbial system. Their deep community structure, dominance patterns, and complex microbial response over space and time serve as an excellent evaluation of next generation sequencing techniques' ability to identify both dominant and subtle details and interactions in these highly diverse systems. These mats produce detectable nighttime fluxes of dissolved hydrogen gas as fermentation byproducts and are of interest as a multi-organism ecological approach to bioenergy production as opposed to a pure culture approach of current commercial algal bioenergy products. Overall, the goal of this study is to track the transcript response of mat-associated organisms as energy passes from sunlight into fixed carbon and nitrogen and subsequently into nighttime fermentation products over the course of a diel cycle. A 24-hr study of Elkhorn Slough, CA microbial mats was conducted in 2011 followed by on-going metagenome and metatranscriptome sequencing by JGI. Field-collected mats were incubated at NASA Ames in artificial seawater media and then sampled over one diel cycle in replicate. Light, temperature, oxygen flux, hydrogen flux, and acetylene reduction assay information were also collected. Randomly selected mats were paired with a molybdate treatment in sulfate-free artificial seawater to inhibit the sulfate reduction pathway and further assess the role of sulfate reducing bacteria in hydrogen metabolism. Four Illumina HiSeq metagenomes have been sequenced and the sequencing of multiple metatranscriptomes are in progress. Using tetranucleotide frequency ESOMs on a single test metagenome, preliminary metagenome bins have been identified for several types of mat-associated organisms. Three ESOM main divisions found in ESOMs correspond to Proteobacteria, Bacteroidetes, and Cyanobacteria. Preliminary results show distinct bins for a clade of sulfur-oxidizing phototrophic purple sulfur Gammaproteobacteria, one bin of several members of the Xanthomonadales clade, several bins from the Flavobacteria and Bacteroides clades, and a bin for a purple non-sulfur Alphaproteobacteria. A high-coverage screen was used to filter out reads of the dominant cyanobacteria *Microcoleus chthonoplastes*. However, other closely related Oscillatoriales cyanobacteria did not appear to cluster into discrete bins. Once bins are resolved for dominant mat members, transcripts will be mapped to sequences in these bins to produce normalized relative abundance profiles of gene expression with time.

Characterization of hemicellulolytic soil microbial communities in disturbed forest stands using stable isotope probing

Hilary Leung* (hilaryl@mail.ubc.ca), Roland Wilhelm, Kendra Mitchell, and William Mohn

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

Forests play a critical role in maintaining biodiversity, mediating atmospheric CO_2 concentrations and global climate. Demand for forest biomass is increasing due to emerging markets for alternative energy and biomaterials alongside traditional uses. A new paradigm for forest management practices is necessary; one which considers the long-term sustainability of soil and considers the microbial processes involved in soil fertility. Large-scale organic matter removal has important implications for soil chemical, biological, and physical properties. Our objective is to identify soil bacterial and fungal populations responsible for carbon cycling, and to characterize the long term effects of organic matter removal on these populations. We have sampled soils from various ecozones across North America in collaboration with a range of forestry researchers conducting the Long-Term Soil Productivity (LTSP) study. By employing stable isotope probing (SIP) of phospholipid fatty-acids (PLFAs) and DNA, we characterized the hemicellulose-degrading bacterial and fungal communities in disturbed forest soil stands. Harvesting organic matter altered the PLFA profiles of these hemicellulolytic microbes in these soil communities in both organic mineral soil horizons but this observation appears to be site dependent. Preliminary SIP-DNA results demonstrate that over a third of our 16s pyrotag libraries classify to *Burkholderia, Massilia, Micrococcineae, Paenibacillus, Caulobacter, Pseudomonas, Sphingobacteriaceae, Pelomonas, and Streptomycineae* generas as the dominant hemicellulolytic bacterial populations. Furthermore, over 1.5% of sequences in our SIP library have been classified under the candidate division TM7, suggesting ecological significance of this uncultured lineage of bacteria. Analysis of hemicellulolytic fungal populations are underway, and results from this study will contribute an improved understanding of microbial communities that drive the carbon cycle, and the effects of timber harvesting on long-term storage of carbon in forest soils.

Use of synthetic biology to redesign plant secondary cell wall deposition

Dominique Loqué* (dloque@lbl.gov)

Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California

The plant cell wall represents a large source of polysaccharides that could be used to substitute for sugar derived from starchy grains, which are currently used to feed and produce biofuels. This lignocellulosic biomass, largely under-utilized, is mainly composed of sugar polymers (cellulose and hemicellulose) embedded very strong aromatic polymer called lignin. Recalcitrant to degradation, lignin inhibits efficient extraction and hydrolysis of the cell wall polysaccharide and prevents cost-effective lignocellulosic-biofuel production. Unfortunately, lignin cannot simply be genetically removed without incurring deleterious consequences on plant productivity. The cost effectiveness of the conversion of the lignocellulosic biomass into sugars is still one of the major components to produce cheap biofuels. Therefore, strategies that can be used to reduce the lignin recalcitrance and that can increase polysaccharide deposition into the cell wall without altering plant growth should be developed. We used synthetic biology to redesign cell wall biosynthesis and deposition without affecting the plant growth. We generated strategies to reduce lignin recalcitrance either by manipulating its spacio-temporal deposition or by manipulating its composition. We also reengineered the control of cell wall

biosynthesis and used it to enhancing cell wall polysaccharides deposition and biomass density. This work conducted by the Joint BioEnergy Institute was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231.

Integrated metagenome and metatranscriptome reveal adaptive ability for sugar degradation, detection and uptake by the cecal microbiota in the leaf-eating flying squirrel (*Petaurista alborufus lena*)

Hsiao-Pei Lu,[1] Yu-bin Wang,[1,2] Chun-Yen Lin,[2] Shiao-Wei Huang,[1] Chih-Hao Hsieh,[3] and Hon-Tsen Yu* (ayu@ntu.edu.tw)[1,4]

[1]Institute of Zoology and Department of Life Science, National Taiwan University, Taipei, Taiwan, Republic of China; [2]Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China; [3]Institute of Oceanography, National Taiwan University, Taipei, Taiwan, Republic of China; [4]Genome and Systems Biology Degree Program, National Taiwan University, Taipei, Taiwan, Republic of China

The diverse gut microbiota function as an effective metabolic system for extracting nutrients and energy from the diet. The gut microbiota of wild herbivores provide a bountiful genetic resource for understanding plant biomass processing, whereas high-throughput sequencing-based studies have been restricted to human and few domesticated animals, such as mouse, cattle, etc. In the present study, we combined metagenomic and metatranscriptomic approaches to investigate the functional characteristics of the cecal microbiota in the leaf-eating flying squirrel (*Petaurista alborufus lena*). Shotgun DNA and RNA from cecal contents of two individuals were sequenced on the Roche 454 GS-FLX Titanium system to evaluate community-wide phylogenetic and functional profiles. Both the metatranscriptome and metagenome were largely dominated by *Firmicutes*, whereas *Proteobacteria*, *Fusobacteria* and *Acidobacteria* had relatively higher expression rates. Functional analyses revealed that the metagenome had relatively more genes about replication, recombination, and defense mechanisms, whereas genes involved in energy production and conversion, carbohydrate transport and metabolism, and cell motility were enriched in the metatranscriptome. Specifically, the genes encoding flagellar components, chemotaxis proteins, and sugar transporters were highly expressed in the cecal microbial community. Furthermore, multiple pathways and diverse glycoside hydrolases involved in the degradation of plant polysaccharides were detected, with abundant beta-glucosidase genes in both the metatranscriptome and metagenome. These results suggest that the cecal microbiota has adapted to the diet and gut environment of the flying squirrel through enhanced ability for sugar degradation, detection and uptake.

Critical advances in amplicon sequencing efficiency

Derek Lundberg* (derek.lundberg@gmail.com),[1] **Scott M. Yourstone*** (scott.yourstone81@gmail.com),[1] Devin Coleman-Derr,[2] Susannah Green Tringe,[2] Jeffrey L. Dangl,[1,3] and Piotr A. Mieczkowski[1,4]

[1]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, North Carolina; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]HHMI-GBMF plant science investigator; [4]Lineberger Comprehensive Cancer Center, Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, North Carolina

Profiling microbial communities through sequencing of 16S rDNA amplicons is a cornerstone technique in metagenomics. We present cost-saving and accuracy-improving innovations for 16S rDNA amplicon sequencing on the Illumina MiSeq platform, many of which are generalizable to other systems. First, base diversity at each sequencing cycle is important for Illumina data quality; we generate diversity with a mix of frame-shifted primers as an alternative to the wasteful but common practice of re-sequencing up to 50% PhiX174 genomic DNA to provide diversity. Second, we produce amplicons in two rounds. Round-1 uses only two cycles to tag each template molecule with unique random nucleotides, while round-2 amplifies the tagged templates and adds a sample barcode. Because sequences sharing the same template tag originated from the same template molecule, consensus building allows *in-silico*reconstruction of the original template molecule, thus dramatically minimizing sequencing error and PCR bias. The ability to mix-and-match a variety of template specific round-1 primers with the same set of universal round-2 barcoding primers provides additional cost-saving opportunities. Finally, metagenomic studies often involve profiling of microbial communities in association with eukaryotic hosts. we use PCR clamps to block amplification of contaminating plastid and mitochondrial DNA sequences from a host plant without biasing the associated microbial community. Together, these techniques improve both the efficiency and accuracy of 16S rDNA amplicon sequencing.

Single molecule, real-time sequencing for base modification detection in eukaryotic organisms: *Coprinopsis cinerea*

Khai Luong* (kluong@pacificbiosciences.com),[1] Tyson A. Clark,[1] Matthew Boitano,[1] Yi Song,[1] Stephen W. Turner,[1] Jonas Korlach,[1] Lukas Chavez,[2] Patricia J. Pukkila,[3] Yun Huang,[2] Virginia K. Hench,[3] William Pastor,[2] Lakshminarayan M. Iyer,[5] Suneet Agarwal,[4] L. Aravind Iyer,[5] and Anjana Rao[2]

[1]Pacific Biosciences, Menlo Park, California; [2]La Jolla Institute for Allergy & Immunology, La Jolla, California; [3]University of North Carolina at Chapel Hill, Department of Biology, Chapel Hill, North Carolina; [4]Harvard Stem Cell Institute, Holyoke Center, Cambridge, Massachusetts; [5]National Center for Biotechnology Information-Computational Biology Branch, Bethesda, Maryland

Single Molecule Real-Time (SMRT®) DNA sequencing provides a wealth of kinetic information beyond the extraction of the primary DNA sequence, and this kinetic information can provide for the direct detection of modified bases present in genomic DNA. This method has been demonstrated for base modification detection in prokaryotes at base and strand resolutions. In eukaryotes, the common base modifications known to exist are the cytosine variants including methyl, hydroxymethyl, formyl and carboxyl forms. Each of these modifications exhibits different signatures in SMRT kinetic data, allowing for unprecedented possibilities to differentiate between them in direct sequencing data. We present early results of directly sequencing different base modifications in eukaryotic genomic DNA using this method.

Metagenomic and metatranscriptomic analysis on TCE-dechlorinating microbial communities enriched under different exogenous cobalamin conditions

Yujie Men* (menyj@berkleley.edu),[1] Julien Tremblay,[2] Emmanuel Prestat,[3] Jacob Bælum,[3,4] Susannah G. Tringe,[2] Janet R. Jansson,[3] and Lisa Alvarez-Cohen[1,3]

[1]Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, California; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [4]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark

Chlorinated ethenes have long been known as common groundwater contaminants in the United States. *Dehalococcoides mccartyi* (Dhc) is, so far, the only anaerobic Bacterium capable of reductively dechlorinating chloroethenes to the innocuous ethene. Dhc exhibits faster dechlorination and cell growth when grown in communities than in isolation, which is possibly due to efficient nutrient exchange between Dhc and other co-existing microorganisms. Corrinoids such as cobalamin are essential cofactors of reductive dehalogenases, but cannot be synthesized by Dhc. Cobalamin is amended to Dhc when grown in isolation, as well as to some enrichment cultures. However, recent studies have shown that Dhc can grow in communities without cobalamin amendment. In order to elucidate the ecological corrinoid interactions within dechlorinating communities, a holistic understanding of functional genes within dechlorinating communities and their interactions is needed. Advances in next generation sequencing technologies allow us to analyze both function and phylogeny within communities in depth. In this study, Illumina sequencing is applied to analyze the metagenomes and metatranscriptomes of two non-methanogenic TCE-dechlorinating communities with (HiTCEB12) and without (HiTCE) exogenous cobalamin. For metagenomic analysis, the 16S rRNA gene reads were screened out using a kmer matching pipeline and the remaining reads were assembled into contiguous sequences (contigs). The 16S rRNA gene composition tracks well with the clone library results identifying the same top five phylogenetic groups: Firmicutes, Chloroflexi, Deltaproteobacteria, Spirocheates and Bacteroidetes. The phylogenetic distribution of protein-encoding genes was analyzed based on BLAST identities and tetranucleotide frequencies. Around 1800 (2.4-2.5% of total) assembled protein-coding contigs belong to Dhc genomes, mostly matching strain 195 with above 90% identity. More than half of the protein-encoding contigs exhibit less than 60% similarity with known genomes, including up-stream corrin ring biosynthesis genes, such as *cbiK*. Genome reconstruction will be performed in order to better understand phylogeny-function relationships. Five metatranscriptomes were also acquired, representing HiTCEB12 and HiTCE cultures at early exponential phase (T1), as well as HiTCE culture at T1, late exponential phase (T2) and stationary phase (T3). The metatranscriptomic sequences match well with the reference metagenomes and will be mapped back according to the reconstructed genomes.

The DOE Systems Biology Knowledgebase: Microbial communities Science Domain

Folker Meyer* (folker@anl.gov), Dylan Chivian, Andreas Wilke, Naryan Desai, Jared Wilkening, Kevin Keegan, William Trimble, Keith Keller, Paramvir Dehal, Robert Cottingham, Sergei Maslov, Rick Stevens, and Adam Arkin

The DOE Systems Biology Knowledgebase (KBase) is a new community resource for predictive biology. It integrates a wide spectrum of data types across the microbial, microbial community, and plant domains, and ties this data into a varied set of powerful computational tools that can analyze and simulate data to predict biological behavior, generate and test hypotheses, design new biological functions, and propose new experiments. The overarching objective is to provide a solid platform that supports predictive biology in a framework that does not require users to learn separate systems to formulate and answer questions spanning a variety of topics in systems biology research. The microbial communities team is integrating both existing and new tools and data into a single, unified framework that is accessible programmatically and through web services. This will allow the construction of sophisticated analysis workflows by facilitating the linkages between data and analysis methods. The standardization, integration and harmonization of diverse data types housed within the KBase and data located on servers maintained by the larger scientific community will allow for a single point of access, ensuring consistency, quality assurance, and quality control checks of data. We have begun by creating KBase data and analysis services that will link our core resources, which will allow clients to access data and analysis methods across these tools without programmatic burdens. New functionality, not currently available in our core tools, is being created within KBase using the programmatic interfaces.

Pangenome: A new perspective to study microbial genomes

Supratim Mukherjee* (supratimmukherjee@lbl.gov),[1] Neha Varghese* (njvarghese@lbl.gov),[1] Aydin Buluc,[2] Nikos C. Kyrpides,[1] Amrita Pati,[1] and Konstantinos Mavromatis[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California

Driven by our quest for knowledge and significant cost-reduction in next generation sequencing technologies in the last decade, microbial genomic information has increased at an astounding rate. This wealth of genomic data is causing a paradigm shift in our approach to study microorganisms. A pangenome, which encompasses core and accessory portions of multiple isolate genomes, represents one such recent development. Comparative pangenome analysis helps uncover important genetic as well as phenotypic variations across member genomes while tracing their evolutionary trajectory. At a time when data storage and analysis is becoming prohibitive, pangenomes come as a boon to provide computational and data management benefits by combining multiple closely related genes into a single pangene. To provide biologically meaningful results, genomes for pangenome creation and analysis, need to be selected based on well-defined metrics. Average nucleotide identity (ANI), which accounts for genome-level similarity between two isolate genomes, was used as a metric to identify pangenome groups. Novel clustering algorithms and paralog finding tools were employed to detect non-redundant genes, which forms the basis for a pangenome. Finally, visualization tools were developed to discover genomic variations between members of a pangenome.

Data integration and analysis processing at EMSL

Angela D. Norbeck* (angela.norbeck@pnnl.gov), David Brown, Kevin Fox, David Cowley, and Samuel Purvine

Environmental and Molecular Sciences Laboratory (EMSL), Richland, Washington

Organisms rarely live in isolation, and it is therefore important that holistic approaches be used to understand the mechanisms of biological adaption to change. Systems biology approaches often incorporate multiple experiment types, and integration of the results is a necessary albeit difficult stage in the discovery process. The Environmental and Molecular Sciences Laboratory (EMSL), located on the

PNNL campus, provides the ability to analyze samples with different capabilities, resulting in data from several experiments, ranging from multi-omics (proteomics, metabolomics, transciptomics) to surface chemistry and imaging. High performance computing allows complex queries to be executed on these data, and visualization efforts to navigate the integrated data types are also in progress. Further, with the recent advent of data exchange with the Joint Genome Institute, joint users are now able to download results from the JGI genome annotation pipeline directly and integrate with experimental data from EMSL. A collaborative scientific research environment has been developed at EMSL, named EMSLHub, which combines data management with analysis pipelines and workflows, along with wiki and blog tools for project tracking. The EMSLHub feeds from an underlying data capture and search tool named MyEMSL, which enables data integration to be performed by users. Features of these developments are presented here, highlighting specific workflows in protein clustering, protegenomics, and targeted re-analysis.

Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes

Robin A. Ohm* (raohm@lbl.gov),[1] Nicolas Feau,[2] Bernard Henrissat,[3] Conrad L. Schoch,[4] Benjamin A. Horwitz,[5] Kerrie W. Barry,[1] Bradford J. Condon,[6] Alex C. Copeland,[1] Braham Dhillon,[2] Fabian Glaser,[7] Cedar N. Hesse,[8] Idit Kosti,[5] Kurt LaButti,[1] Erika A. Lindquist,[1] Susan Lucas,[1] Asaf A. Salamov,[1] Rosie E. Bradshaw,[9] Lynda Ciuffetti,[8] Richard C. Hamelin,[2,10] Gert H. J. Kema,[11] Christopher Lawrence,[12] James A. Scott,[13] Joseph W. Spatafora,[8] B. Gillian Turgeon,[6] Pierre J.G.M. de Wit,[14] Shaobin Zhong,[15] Stephen B. Goodwin,[16] and Igor V. Grigoriev[1*]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of British Columbia, Vancouver, British Columbia, Canada; [3]Aix-Marseille Université, Marseille, France; [4]National Institutes of Health/National Library of Medicine/National Center for Biotechnology Information, Bethesda, Maryland; [5]Department of Biology, Technion, Israel Institute of Technology, Haifa, Israel; [6]Cornell University, Ithaca, New York; [7]Bioinformatics Knowledge Unit, Technion, Israel Institute of Technology, Haifa, Israel; [8]Oregon State University, Corvallis, Oregon; [9]Massey University, Palmerston North, New Zealand; [10]Natural Resources Canada, Ste-Foy, Quebec, Canada; [11]Plant Research International, Waginingen, The Netherlands; [12]Virginia Bioinformatics Institute & Department of Biological Sciences, Blacksburg, Virginia; [13]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [14]Wageningen University, Wageningen, The Netherlands; [15]North Dakota State University, Fargo, North Dakota; [16]Purdue University, West Lafayette, Louisiana

The class *Dothideomycetes* is one of the largest groups of fungi with a high level of ecological diversity including many plant pathogens infecting a broad range of hosts. Here, we compare genome features of 18 members of this class, including 6 necrotrophs, 9 (hemi)biotrophs and 3 saprotrophs, to analyze genome structure, evolution, and the diverse strategies of pathogenesis. The *Dothideomycetes* most likely evolved from a common ancestor more than 280 million years ago. The 18 genome sequences differ dramatically in size due to variation in repetitive content, but show much less variation in number of (core) genes. Gene order appears to have been rearranged mostly within chromosomal boundaries by multiple inversions, in extant genomes frequently demarcated by adjacent simple repeats. Several *Dothideomycetes* contain one or more gene-poor, transposable element (TE)-rich putatively dispensable chromosomes of unknown function. The 18 *Dothideomycetes* offer an extensive catalogue of genes involved in cellulose degradation, proteolysis, secondary metabolism, and cysteine-rich small secreted proteins. Ancestors of the two major orders of plant pathogens in the *Dothideomycetes*, the *Capnodiales* and *Pleosporales*, may have had different modes of pathogenesis, with the former having fewer of these genes than the latter. Many of these genes are enriched in proximity to transposable elements, suggesting faster evolution because of the effects of repeat induced point (RIP) mutations. A syntenic block of genes, including oxidoreductases, is conserved in most *Dothideomycetes* and upregulated during infection in *L. maculans*, suggesting a possible function in response to oxidative stress.

Functional genomics of lignocellulose degradation in the Basidiomycete white rot *Schizophyllum commune*

Robin A. Ohm* (raohm@lbl.gov),[1] Martin Tegelaar,[2] Bernard Henrissat,[3] Heather M. Brewer,[4] Samuel O. Purvine,[4] Scott Baker,[4] Han A. B. Wösten,[2] Igor V. Grigoriev,[1] and Luis G. Lugones[2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Utrecht University, Utrecht, The Netherlands; [33]Aix-Marseille Université, Marseille, France; [4]Pacific Northwest National Laboratory, Richland, Washington

White and brown rot fungi are among the most important wood decayers in nature. Although more than 50 genomes of Basidiomycete white and brown rots have been sequenced by the Joint Genome Institute, there is still a lot to learn about how these fungi degrade the tough polymers present in wood. In particular, very little is known about how these fungi regulate the expression of genes involved in lignocellulose degradation. In Ascomycetes, several conserved transcription factors involved in regulation of complex carbon source degradation have been identified, but there are no homologs of these in Basidiomycetes. Few Basidiomycete white or brown rots are genetically amenable, hindering a functional genomics approach to the study of lignocellulose degradation. A notable exception is *Schizophyllum commune*, for which numerous genetic tools are available. *S. commune* was grown on several carbon sources (glucose, cellulose, and beech wood) and gene expression was analyzed. Numerous genes are strongly up-regulated on the complex carbon sources, compared to on glucose. As expected, many of these encode CAZymes (notably glycoside hydrolase family 61, or GH61) and FOLymes, but also several well conserved proteins with unknown function. Interestingly, five transcription factor genes are up-regulated during growth on complex carbon sources, suggesting they may be involved in regulating this process. These transcription factors are highly conserved in Basidiomycetes, but not in Ascomycetes. The two laccase genes of *S. commune* are very lowly expressed on complex carbon sources, suggesting that their function in lignocellulose degradation is limited. The secretome was analyzed during growth on glucose and cellulose, using mass spectrometry. In total, 672 proteins were identified, and 84 of those were especially abundant. The majority consists of CAZymes and FOLymes, and has a secretion signal. Members of the GH61 CAZyme family are strongly over-represented. A promoter analysis of a strongly up-regulated member of the GH61 family reveals that the 700 bp region upstream of the translation start site is capable of inducing dTomato fluorescence on cellulose, but not on glucose. This is in agreement with both the transcriptomics and proteomics data. Comparison of promoters of other up-regulated members of the GH61 family reveals a conserved putative transcription factor binding site.

Haplotype assembly refinement and improvement

Christian Olsen* (Christian@biomatters.com),[1] Kashef Qaadri,[1], Matthew Shoa-Azar,[2] Joan Wong,[2] Trung Nguyen,[2] George Rudenko,[2] Tina Huynh,[2] Aravind Somanchi,[2] Jeff Moseley,[2] Scott Franklin,[2] and **Shane Brubaker**[2]

[1]Biomatters, Inc., Newark, New Jersey; [2]Solazyme, South San Francisco, California

Haplotyping, the resolution of two distinct alleles of an organism, is an important and challenging problem in BioInformatics. In particular, the phasing of SNPs (correct assignment of alleles) across long distances is very challenging given today's Next Generation Sequencing (NGS) technologies. We have used Geneious to analyze reads from several technology platforms, including 454, Illumina, and PacBio. Key features of the software including read mapping, read visualization and mate pair analysis, and editing and polymorphism analysis were used to support the workflow. We refined the workflow with an emphasis on speed and accuracy. We have been able to manually haplotype a highly polymorphic, diploid organism over arbitrarily long distances. The process is far faster and more accurate using PacBio reads, although we found a second QC pass with Illumina data to be highly valuable. We verified haplotyping accuracy using known bacterial artificial chromosome (BAC) sequences. In addition, we have explored attempts to perform automated diploid assembly with correct haplotyping phasing, with mixed results. We use the improved Geneious 6.0 *de novo* assembler. We discuss these results here and make suggestions for further improvements to assembly algorithms to help support haplotyped assembly.

Comparison of three Cre-loxP based paired-end library construction methods

Ze Peng* (zpeng@lbl.gov), Nandita Nath, Andrew Tritt, Shoudan Liang, James Han, Chia-Lin Wei, Len Pennacchio, and Feng Chen

DOE Joint Genome Institute, Walnut Creek, California

Paired-end library sequencing has been proven useful in scaffold construction during *de novo* whole genome shotgun assembly. The ability of generating mate pairs with > 8 Kb insert sizes is especially important for genomes containing long repeats. To make mate paired libraries for next generation sequencing, DNA fragments need to be circularized to bring the ends together. There are several methods that can be used for DNA circulation, namely ligation, hybridization and Cre-LoxP recombination. With higher circularization efficiency with large insert DNA fragments, Cre-LoxP recombination method generally has been used for constructing >8 kb insert size paired-end libraries. Second fragmentation step is also crucial for maintaining high library complexity and uniform genome coverage. Here we will describe the following three fragmentation methods: restriction enzyme digestion, random shearing and nick translation. We will present the comparison results for these three methods. Our data showed that all three methods are able to generate paired-end libraries with greater than 20 kb insert. Advantages and disadvantages of these three methods will be discussed as well.

Rumen systems microbiology: Towards a functional systems-level understanding of the microbial community and the biomass degradation process in the cow rumen

Hailan Piao*,[1] Stephanie Malfatti,[2] Alexander Sczyrba,[3] Julien Tremblay,[2] Robin Ohm,[2] Kanwar Singh,[2] Fernanda Haffner,[4] Stefan Bauer,[4] David Culley,[5] Kenneth Bruno,[5] Kerrie Barry,[2] Feng Chen,[2] Scott Baker,[6] Susannah Tringe,[2] Igor Grigoriev,[2] Roderick Mackie,[7] and **Matthias Hess** (matthias.hess@tricity.wsu.edu)[1,5,6]

[1]Washington State University, Richland, Washington; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]University of Bielefeld, Bielefeld, Germany; [4]University of California at Berkeley, Energy Biosciences Institute, Berkeley, California; [5]Pacific Northwest National Laboratory, Chemical and Biological Process Development Group, Richland, Washington; [6]Pacific Northwest National Laboratory, Environmental Science Molecular Laboratory, Richland, Washington; [7]University of Illinois, Urbana-Champaign, Illinois

The microbial community that inhabits the cow's rumen is the most efficient biomass degrading system on Earth. Although the rumen microbiome has attracted a lot of attention over the last decades and microbiologists have identified and isolated several lignocellulolytic organisms and enzymes from this environment, it is still not possible to synthesize a "true rumen microbiome" in the laboratory. This explains why it is still not well understood how microorganisms and their molecular components affect and respond to physiological changes in the rumen. Recent advances in the field of diverse *-omics* techniques and the development of affordable computer hardware and computational software have made it possible to enhance our multi-level understanding of microbial systems - and how they interact with their environment. With these tools we are now adequately equipped to tackle problems of a dimension too complex for traditional molecular techniques. Here we will present the first results of a multi-institutional project in which a suite of cultivation-independent techniques (i.e. phylogenomics, metagenomics, metatranscriptomics, and metaproteomics) are utilized to enhance our understanding of the community dynamics and the microbial processes that occur during biomass degradation in the anaerobic rumen of a cow. The microbial community that inhabits the cow rumen is composed of Archaea, Bacteria and Eukarya and is well known for its biomass-degrading ability. In order to understand this ecosystem at the whole-systems level it is important to monitor the dynamics of the individual community members. To obtain insights into the ecology of the rumen community and its dynamics during biomass-degradation, we amplified a short region of the 16S rRNA gene and the ITS2 region from the prokaryotic and eukaryotic population that colonized biomass during rumen incubation respectively. Amplicons were generated from rumen-incubated biomass at different time points using Roche’s 454 Titanium and Illumina’s MiSeq platform. We generated a total of ~11 million sequences with an average read length of ~500bp (Roche) and ~240bp (Illumina) amounting to a total of >2.7 Gbp of sequence information. Succeeding sequence analysis revealed a fungal community of low complexity, with two phyla as the dominant players. Members of the phylum Neocallimastigomycota were absent on the pre-incubated biomass and appeared to colonize both corn stover and switchgrass throughout the incubation process. Members of the phylum Ascomycota were less dominant (<1%), but seemed to play a role in the later phases of the biomass-degradation process. Analysis of the prokaryotic community revealed ~1,000 OTUs, with members belonging to two phylogenetic classes as the dominant players, and a separation of the prokaryotic population into early and late colonizers. In summary, results presented here demonstrate that rumen prokaryotes and fungi consistently colonize lignocellulosic substrates and that certain populations contribute during different stages of the degradation process of recalcitrant biomass in the rumen ecosystem.

A custom database for functional annotation of soil omics datasets using Hidden Markov Models

Emmanuel Prestat* (EPrestat@lbl.gov),[1,2] Maude M. David,[1] Konstantinos Mavromatis,[1,3] and Janet R. Jansson[1,3]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Kansas State University, Manhattan, Kansas; [3]DOE Joint Genome Institute, Walnut Creek, California

A current bottleneck in meta-omics analyses of soil communities is the lack of tools to accurately assess functional information in an acceptable computation time. To address this need, we aimed to build a comprehensive manually curated and validated database for screening of omics datasets. We focused on specific biochemical functions and pathways of importance for soil microbial ecology, by manually selecting and organizing functional gene information into categories here called "Maudules". KEGG orthologs sets (KOs) were retrieved to fit within a hierarchical organization of specific pathways of interest (such as denitrification, methanogenesis, etc.). The reduced size of the database (compared to non- specific sequence databases) was a first step towards significant improvement of the duration and specificity of similarity searches. In addition, to improve upon the speed and sensitivity of conventional BLAST searches, we turned each KO set into Markov models by fetching their corresponding pfam profiles. This step generated a sizeable number of conflicts (several pfam per KO or the opposite) automatically resolved by functional assignments to KO. For the few remaining unresolved assignations, the corresponding set of sequences was manually split according to the topology of their phylogenetic trees. At this point the HMM were re-trained from the new pool of sequences. To assess the Hidden Markov Model ability to classify sequences into functions in terms of sensitivity and specificity, we have screened the whole Kegg proteins database with our models and compared our KO classification and the "genuine" Kegg annotation. The resulting product is the first soil-specific HMM database, allowing us to bin increasingly large "omics" datasets such as metagenomes, metatranscriptomes and metaproteomes from soil samples into functions. All the annotation methods developed in this project will be made available for the scientific community through KBase.

Comparative analysis of 35 basidiomycete genomes reveals diversity and uniqueness of the phylum

Robert Riley* (rwriley@lbl.gov),[1] Asaf Salamov,[1] Robert Otillar,[1] Kirsten Fagnan,[1] Bastien Boussau,[2] Daren Brown,[3] Bernard Henrissat,[4] Anthony Levasseur,[4] Benjamin Held,[5] Laszlo Nagy,[6] Dimitris Floudas,[6] Emmanuelle Morin,[7] Gerard Manning,[8] Scott Baker,[3] Francis Martin,[6] Robert Blanchette,[9] David Hibbett,[6] and Igor V. Grigoriev[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of California at Berkeley, Berkeley, California; [3]Pacific Northwest National Laboratory, Richland, Washington; [4]Aix-Marseille Université, Marseille, France; [5]University of Minnesota, St. Paul, Minnesota; [6]Clark University, Worcester, Massachusetts; [7]INRA, Champenoux, France; [8]Salk Institute for Biological Sciences, San Diego, California; [9]University of Minnesota, St. Paul, Minnesota

Fungi of the phylum Basidiomycota (basidiomycetes), make up some 37% of the described fungi, and are important in forestry, agriculture, medicine, and bioenergy. This diverse phylum includes symbionts, pathogens, and saprobes including wood decaying fungi. To better understand the diversity of this phylum we compared the genomes of 35 basidiomycete fungi including 6 newly sequenced genomes. The genomes of basidiomycetes span extremes of genome size, gene number, and repeat content. A phylogenetic tree of Basidiomycota was generated using the Phyldog software, which uses all available protein sequence data to simultaneously infer gene and species trees. Analysis of core genes reveals that some 48% of basidiomycete proteins are unique to the phylum with nearly half of those (22%) comprising proteins found in only one organism. Phylogenetic patterns of plant biomass-degrading genes suggest a continuum rather than a sharp dichotomy between the white rot and brown rot modes of wood decay among the members of Agaricomycotina subphylum. There is a correlation of the profile of certain gene families to nutritional mode in Agaricomycotina. Based on phylogenetically-informed PCA analysis of such profiles, we predict that that *Botryobasidium botryosum* and *Jaapia argillacea* have properties similar to white rot species, although neither has liginolytic class II fungal peroxidases. Furthermore, we find that both fungi exhibit wood decay with white rot-like characteristics in growth assays. Analysis of the rate of discovery of proteins with no or few homologs suggests the high value of continued sequencing of basidiomycete fungi.

Single-cell genomics at the JGI – Illuminating microbial dark matter

Christian Rinke* (crinke@lbl.gov),[1] Patrick Schwientek,[1] Alexander Sczyrba,[1,2] Natalia N. Ivanova,[1] Iain J. Anderson,[1,‡] Jan-Fang Cheng,[1] Aaron Darling,[3,4] Stephanie Malfatti,[1] Brandon K. Swan,[5] Esther A. Gies,[6] Jeremy A. Dodsworth,[7] Brian P. Hedlund,[7] George Tsiamis,[8] Stefan M. Sievert,[9] Wen-Tso Liu,[10] Jonathan A. Eisen,[3] Steven Hallam,[6] Nikos C. Kyrpides,[1] Ramunas Stepanauskas,[5] Edward M. Rubin,[1] Philip Hugenholtz,[11] and Tanja Woyke[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Center for Biotechnology, Bielefeld University, Bielefeld, Germany; [3]Department of Evolution and Ecology, University of California Davis, Davis, California; [4]ithree institute, University of Technology Sydney, Sydney, Australia; [5]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine; [6]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada; [7]School of Life Sciences, University of Nevada, Las Vegas, Nevada; [8]Department of Environmental and Natural Resources Management, University of Western Greece, Agrinio, Greece; [9]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts; [10]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois; [11]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences and Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Australia. [‡]Deceased.

Genome sequencing enhances our understanding of the biological world by providing blueprints for the evolutionary and functional diversity that shapes the biosphere. Currently available microbial genomes, however, are of limited phylogenetic breadth due to our historical inability to cultivate most microorganisms in the laboratory. Applying single-cell genomics, we targeted and sequenced 201 uncultivated archaeal and bacterial cells from nine diverse habitats belonging to 28 major uncharted branches of the tree of life, so-called “microbial dark matter”. With this additional genomic information, we could resolve numerous intra- and inter-phylum level relationships and propose a new archaeal superphylum. We discovered unexpected metabolic features that extend our understanding of biology and challenge established boundaries between the three domains of life. These include a novel amino acid usage for the opal stop codon, an archaeal-type purine synthesis in Bacteria and complete bacterial-like sigma factors in Archaea. The single-cell genomes also served to phylogenetically anchor up to 20% of metagenomic reads in some habitats, facilitating organism-level interpretation of ecosystem function. This study greatly expands the genomic representation of the tree of life and provides a systematic step forward to a better understanding of microbial evolution on our planet.

A metagenomic approach for identification of novel enzymes from the Red Sea Atlantis II brine pool

Ahmed Said, **Mohamed Ghazy*** (mghazy@aucegypt.edu), Mohamed Maged, Amged Ouf, Ari Ferrira, Rania Siam, and Hamza El Dorry

Department of Biology, The American University in Cairo, New Cairo, Egypt

The Atlantis II brine pool, of about 60 km^2, is located in the central part of the Red Sea (21°21' N, 38°04' E) at a depth of 2000 to 2200 meters below the sea. The lower part of this pool, the lower convective layer (ATII-LCL), is characterized by extreme harsh environmental conditions; the temperature reaches 68.2°C, salinity of 270 psu, it is almost anoxic, and contains a very high concentration of heavy metals. Microbial communities inhabiting this harsh environment are expected to have enzymes and proteins that are catalytically adapted to extreme salinity, high temperature, and high concentration of heavy metals. Enzymes from microorganisms that are adapted to function at high temperature, extreme salinity or resistant to high concentration of heavy metals are described in the literatures. However, enzymes that are collectively adapted to these three extreme environmental factors have not been described and characterized yet. In order to shed light on the structural-functional relationship of enzymes from the microbial community inhabiting this unique environment, and to understand how enzymes adapt collectively to these chemical and physical biotic conditions, we have carried out metagenomic analysis of DNA isolated from microbial community collected from the ATII-LCL. In this regards, we have established a 454-pyrosequencing metagenomic dataset, and a functional screening methods for different hydrolyses. In addition, using our dataset, we identified genes and operons involved in heavy metals detoxification processes. One of the operon identified in our dataset is for mercury detoxification, including mercuric reductase (MerA), a key component for the detoxification system found in many bacteria adapted to such environment. Mercuric reductase catalyzes the reduction of inorganic mercuric ions (Hg^{+2}) to elemental mercury (Hg^0), which is volatile, and is consequently removed from the cell. This metagenome-derived gene, and its orthologue from an uncultured soil microorganism were expressed in *E. coli*, these proteins were purified, and the enzymatic properties were compared. *E. coli* ATII-LCL *MerA* transformants can grow on high concentration of $HgCl_2$ while the soil transformants was sensitive to low concentration of mercury. Unlike the soil MerA enzyme, the purified ATII-LCL MerA enzyme was activated by sodium chloride, and remained active at a temperature of up to 70 ^oC. Interestingly, differences in the enzymatic properties of the two orthologs were attributed to less than 9 % differences in amino acid compositions, from which 70% are acidic amino acids substitutions. The work shows that the enzymatic activities of both proteins reflect adaptation to their environmental conditions. Site directed mutagenesis pinpointed specific residues that are required for extreme halophilicity, resistance to high concentration of mercuric chloride, and stability to high temperature.

This work was performed in collaboration with King Abdulah University for Science and Technology (KAUST).

Comparative reannotation of 21 Aspergillus genomes

A. Salamov* (aasalamov@lbl.gov), R. Riley, A. Kuo, and I. Grigoriev

DOE Joint Genome Institute, Walnut Creek, California

We used comparative gene modeling to reannotate 21 Aspergillus genomes. Initial automatic annotation of individual genomes may contain some errors of different nature, like, f.e missing some

genes, non-correct exon-intron structures, 'chimeras', which fuse 2 or more real genes or alternatively splitting some real genes into 2 or more models. The main premise behind the comparative modeling approach is that for closely related genomes most orthologous families have the same conserved gene structure. The algoritm maps all gene models predicted in all individual Aspergillus genomes to each genomes and for each locus selects among the potentially many competing models, the one which most closely resembles the ortologous genes from other genomes. This procedure is iterated until no change in gene models will be observed. For Aspergillus genomes we in total predicted 4503 new gene models (~2% per genome), supported by comparative analysis, additionally correcting ~18% of old gene models. This resulted in total of 4065 more genes with annotated PFAM domains(~3% increase per genome). Analysis of few genomes with EST/transcriptomics data shows that new annotation sets also have a higher number of EST-supported splice sites at exon-intron boundaries.

Phasing variants in poplar trees using a hybrid of short and long read technologies

Wendy Schackwitz* (wsschackwitz@lbl.gov),[1] Joel Martin,[1] Anna Lipzen,[1] Len Pennacchio,[1] and Gerald Tuskan[2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee

Determining the phase of variants is useful for detecting regions of recombination, imputing missing data, correcting genotyping errors, and ascertaining the independence of multiple neighboring variants in polyploid genomes. One approach for directly phasing variants is utilizing sequencing technology that generates long contiguous reads from a single DNA molecule/chromosome. Previously this was accomplished by 1st generation Sanger sequencing and while this method is effective; it is expensive, time consuming, and the reads are limited to around 1kb in length, constraining the size of the region that can be phased and thereby of limited utility for most eukaryotic genomes. With cost effective 2nd generation sequencing technologies on the market the contiguous region of the genome that can be phased is restricted by the short read lengths obtained. Strategies with varying inserts lengths and paired end reads can help but the construction of multiple libraries with multiple insert sizes significantly increases the cost and time to generate individual chromosome haplotypes. 3rd generation sequencing technologies generate long reads potentially allowing the direct phasing of DNA blocks, however since this technology is not as high throughput as 2nd generation and has a higher error rate, it is challenging to generate the sequence depth necessary for accurate variant calling. By employing a combination of 2nd and 3rd generation sequencing technologies, we can benefit from the accuracy and depth of sequence generated from 2nd generation while also having the long reads of the 3rd generation needed to phase distant variants. Here, we will present the results of using both 2nd and 3rd generation sequencing technologies to determine the haplotype of genomes for individual poplar trees of relevance to biofuel development.

Roles of genotype-by-environment interactions in shaping the root-associated microbiome of *Populus*

Christopher W. Schadt* (schadtcw@ornl.gov),**[1,2]** Migun Shakya,[1,2] Michael Robeson,[1] Neil Gottel,[1] Hector Castro,[1] Zamin Yang,[1] Marilyn Kerley,[1] Gregory Bonito,[3] Dale Pelletier,[1,2] Susanna Tringe,[4] Stephanie Malfatti,[4] Kanwar Singh,[4] Tijana Glavina del Rio,[4] Sagar Utturkar,[1,2] Tatiana Karpinets,[1] Jesse Labbe,[1] Wellington Muchero,[1] Steven D. Brown,[1,2] Francis Martin,[5] Mircea Podar,[1,2] Rytas Vilgalys,[3] Mitchel J. Doktycz,[1,2] and Gerald Tuskan[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]University of Tennessee, Knoxville, Tennessee; [3]Duke University, Durham, North Carolina; [4] DOE Joint Genome Institute, Walnut Creek, California; [5] Institut National de la Recherche Agronomique, Nancy Université, France

Populus trees represent a genetically diverse, ecologically widespread, perennial woody genus, that have potential as cellulosic feedstocks for biofuels and contain the first tree species to have a full genome sequence. These trees are also host to a wide variety of symbiotic microbial associations within their roots and rhizosphere, thus may serve as ideal models to study the breadth and mechanisms of interactions between plants and microorganisms. However, most of our knowledge of *Populus* microbial associations to date comes from greenhouse and plantation-based trees; there have been few efforts to comprehensively describe microbial communities of mature natural populations of *Populus*. We have compared root endophyte and rhizosphere samples collected from two dozen sites within two watershed populations of *Populus deltoides* in eastern Tennessee and western North Carolina over multiple seasons. 454 pyrosequencing was applied to survey and quantify the microbial community associated with *P. deltoides*, using primers targeting the bacterial 16S rRNA gene and the fungal 28S rRNA gene. Genetic relatedness among the *Populus* trees was evaluated using 20 SSR markers chosen for distribution across all 19 linkage groups of the *Populus* genetic map. Soil physical, chemical and nutrient status, as well as tree growth and age characteristics were also evaluated. Root endosphere and rhizosphere communities were found to be composed of distinct assemblages of bacteria and fungi with largely non-overlapping OTU distributions. Within these distinct endophyte and rhizosphere habitats, community structure is influenced by soil characteristics, watershed origin and seasonal changes in some cases; while observed host genotype effects have been minimal at the relatively course resolution of the SSR markers. Our current CSP project with JGI is exploring plant genotypic influences on *Populus trichocarpa* soil, rhizosphere and endosphere communities using a well studied collection of over 1000 resequenced genotypes arrayed in common gardens in the Pacific Northwest. iTAGs libraries are currently being sequenced for over 550 samples obtained from two common garden sites in Oregon last spring. Initial results suggest similar distributions of organisms at the Phylum level as were observed previously in eastern *P. deltoides* populations and include many similar dominant OTUs of *Pseudomonas* and *Streptomyces*-like organisms in the root endosphere. We have isolated cultures of over a 1500 bacteria and fungi from these environments representing many of these dominant community members *in situ* and many of these isolates show distinct growth-promoting phenotypes with *Populus* in greenhouse resynthesis experiments. Fourty three of the bacterial isolates and 5 of the fungal isolates now have draft genome sequence available, including 20 *Pseudomonas fluorescens* isolates that vary widely in genetic makeup and host association properties. Our findings suggest that the characteristics of the *Populus* root/soil environment may represent a relatively strong selective force in shaping endophyte and rhizosphere microbial communities. Forthcoming work in collaboration with JGI will explore the genetic basis of these associations within the resequenced *P. trichocarpa* common garden populations and metagenomic analyses of selected genotype/phenotype associations. More information about these efforts can be found on our project web portal @ http://pmi.ornl.gov.

The PhyloFacts FAT-CAT Web server: Functional (and taxonomic) annotation of (meta)genomes across the Tree of Life

Kimmen Sjölander* (kimmen@berkeley.edu),[1,2] Cyrus Afrasiabi,[1] Bushra Samad,[1] and Dave Dineen[1]

[1]QB3 Institute, University of California at Berkeley, Berkeley, California; [2]Department of Bioengineering and Department of Plant and Microbial Biology, University of California at Berkeley, Berkeley, California

The PhyloFacts FAT-CAT web server is a new web server in the PhyloFacts suite of phylogenomic tools, providing highly precise phylogenetic placement of protein sequences within gene trees in the PhyloFacts database. The PhyloFacts database is unique among phylogenomic resources of gene families in both taxonomic breadth -- more than 7M proteins from 99K unique taxa (including strains) from Bacteria, Archaea, Eukarya and viruses – and functional annotation capabilities, integrating experimental data and annotations from numerous resources over >93K trees for Pfam domains and multi-domain architectures. FAT-CAT makes use of hidden Markov models (HMMs) placed at all nodes in PhyloFacts family trees for phylogenetic classification of user-submitted sequences; when the top-scoring HMM is located within a clade of orthologs, the FAT-CAT classification provides highly specific functional subclassification and ortholog identification. We provide four pipeline parameter presets to handle different sequence types, including partial sequences and proteins containing promiscuous domains; users can also modify individual parameters. PhyloFacts trees matching the query can be viewed interactively online using the PhyloScope Javascript tree viewer and are hyperlinked to various external databases. Users can also explore remote evolutionary relationships to obtain additional clues to function and/or structure. FAT-CAT is computationally intensive and can require one or more hours to complete; a streamlined version of the FAT-CAT pipeline, called FAST-CAT, is provided for users preferring a rapid turn-around. The FAT-CAT webserver is available at http://phylogenomics.berkeley.edu/phylofacts/fatcat/. The FAT-CAT webserver and the PhyloFacts database are funded by a grant from the DOE Systems Biology Knowledgebase.

Molecular systems synecology of TCE-dechlorinating microbial communities using stable isotope probing combined with Illumina 16S rRNA-targeted and shotgun sequencing

B. Stenuit * (bstenuit@berkeley.edu),[1] J. Tremblay,[2] X. Mao,[1] Y. Men,[1] S.G. Tringe,[2] and L. Alvarez-Cohen[1]

[1] Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, California; [2] DOE Joint Genome Institute, Walnut Creek, California

The optimization of *Dehalococcoides*-based bioremediation technologies to treat trichloroethene (TCE)-contaminated groundwater and completely reduce TCE to non-toxic ethene requires comprehensive predictions of interspecies electron, energy and metabolite transfers that shape the structural and functional robustness of TCE-dechlorinating microbial communities. Molecular systems synecology of such syntrophic communities provides systems-level understanding of community structure and function as well as the interactions between community members and their influence on the overall behavior of the system using emerging high-throughput molecular biology tools (e.g., high-depth-resolution, next-generation sequencing (NGS) technologies). In this work, a systems-level functional profiling of three dissimilar TCE-dechlorinating microbial communities is described using a combination of nucleic acid stable isotope probing (NA-SIP) and illumina-based sequencing techniques. NA-SIP was applied to identify microorganisms supporting TCE dechlorination in anaerobic *Dehalococcoides*-containing mixed cultures that have been functionally stable (generating ethene from TCE) for several years using lactate as electron donor. After target microbial communities were subcultured with 5 mM $^{13}C_3$-labeled lactate, ^{13}C-labeled RNA and DNA were isolated by isopycnic ultracentrifugation and multiplex illumina 16S rRNA and shotgun libraries were constructed from heavy and light SIP fractions. Barcoded and staggered length primers targeting the 16S rRNA hypervariable region V4 were used to generate amplicons of about 415 bp (illumina-based 16S-tags (*iTags*)). Illumina shotgun libraries (size of ~370 bp) were constructed on the IntegenX Apollo 324 System (Functional Genomics Laboratory, UC Berkeley) using semi-custom protocols based on standard DNA library preparation for illumina-compatible sequencing. 150-nucleotide paired-end multiplex sequencing was performed with the HiSeq™ 2000 platform followed by *in silico* assembly of paired-end reads and phylogenetic and functional analysis of metagenomic data. The combination of NA-SIP and illumina sequencing techniques is an extremely sensitive approach to decode function-targeted metagenomes characterized by lower complexity and higher resolution/coverage for specific active functional guilds (including low-abundance community members such as syntrophs). This work demonstrates the valuable combination of SIP and illumina sequencing for a fundamental understanding of (i) syntrophic lifestyles governing anaerobic dechlorinating communities and (ii) mutually beneficial cross-feeding interactions with the identification of specific metabolically active functional guilds that support *Dehalococcoides* activity.

Diversity of picophytoplankton along a physico-chemical gradient in the Northeastern Pacific

Sebastian Sudek* (ssudek@mbari.org),[1] R. Craig Everroad,[2] Alyssa Gehman,[1] and Alexandra Z. Worden[1]

[1]Monterey Bay Aquarium Research Institute, Moss Landing, California; [2]Exobiology Branch, NASA Ames Research Center, Moffett Field, California

Marine phytoplankton are responsible for half of the global uptake of CO_2 via photosynthesis. Picophytoplankton (diameter <2-3 um) dominate open ocean environments and are composed of three major groups: *Prochlorococcus*, *Synechococcus* and diverse picoeukaryotes. We investigated the composition of picophytoplankton communities in a Pacific Ocean eastern boundary current system. Samples were taken along a gradient from coastal waters to relatively oligotrophic waters 800 km off-shore, traversing a mesotrophic, upwelling influenced region and the California Current en route. Picophytoplankton groups were enumerated by flow cytometry and formed distinct ratios representing different biogeochemical zones along the gradient. To investigate diversity, we used Sanger (full-length) and barcoded 454-titanium (V1-2 region) sequences of the 16S rRNA gene. Samples were analysed from sunlit regions of the water column representing the surface and either the base of the mixed layer or the deep chlorophyll maximum (when present, ~ 80 m). Initial analyses using the Ribosomal Database Project Classifier (RDPC) tool identified 600-14,000 cyanobacterial sequences per sample, with the fewest in the coastal and the largest number at the open-ocean station. This reflects the abundance of cyanobacteria among the pool of 16S sequences at each station. General sequence abundances corresponded to population trends seen by flow cytometry. *Prochlorococcus* and *Synechococcus* dominated (>90% of total picocyanobacteria) open-ocean and coastal zones, respectively. In the mesotrophic region differences in relative *Prochlorococcus* to *Synechococcus* sequence abundances were less pronounced. Although *Prochlorococcus* and *Synechococcus* are considered genera, they each harbour tremendous diversity which has been studied extensively at genomic and phylogenetic levels. Using 786 full-length cyanobacterial sequences (Sanger) generated from the study site we refined existing 16S rDNA phylogenies and discovered two new *Synechococcus* clades, which formed a significant proportion of sequences at the mesotrophic station. We then used two approaches to assign barcoded sequences to defined clades. First, we designed *in-silico* probes to discriminate eight *Prochlorococcus* and thirteen *Synechococcus* clades based on the alignment of full-length sequences used in phylogenetic reconstructions. The probes identified ~85% and ~70% of RDPC-recognized cyanobacterial sequences in surface and deep samples, respectively. Secondly, we applied a newly developed bioinformatics pipeline (Phyloassigner[1]) that uses maximum likelihood methods to place sequences on the reference tree without prior binning into operational taxonomical units (OTUs), which can obscure phylogenetic information. The pipeline also assists in identifying potentially new clades by including basal nodes on the tree into the analysis. An investigation of these basal sequences is ongoing. Additionally, the two methods agreed well and both were able to distinguish between very closely related clades. Details of the methods and analysis results will be presented. The approaches used should facilitate high-throughput studies on picophytoplankton diversity and screening of 'omic samples for unique taxa. Moreover, our results show that novel picocyanobaterial groups are still being identified even in well-studied marine systems.

[1] *Vergin KL, Beszteri B, Monier A, Cameron Thrash J, Temperton B, Treusch AH, Kilpert F, Worden AZ, Giovannoni SJISME J doi:10.1038/ismej.2013.32*

Genome organization and gene expression in *Miscanthus*

Kankshita Swaminathan* (kank@illinois.edu),[1] **Therese Mitros*** (tmitros@gmail.com)**,**[3,4] Adam Barling,[1,2] Won Byoung Chae,[1,2,7] Brandon T. James,[1,2] Jessica Kirkpatrick,[1,2] Liang Xie,[1] Katarzyna Glowacka,[1,5] Magdy Alabady,[1] Stanislaw Jezowski,[5] John A. Juvik,[1,2] Matthew Hudson,[1,2] Daniel S. Rokhsar,[3,4,6] and Stephen P. Moose[1,2]

[1]Energy Biosciences Institute, Institute for Genomic Biology, University of Illinois, Urbana, Illinois; [2]Crop Sciences, Edward R. Madigan Laboratory, University of Illinois, Urbana, Illinois; [3]Energy Biosciences Institute, University of California at Berkeley, Berkeley California; [4]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California; [5]Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland; [6]DOE Joint Genome Institute, Walnut Creek, California; [7]National Institute of Horticultural & Herbal Science, Rural Development Administration, Suwon, Republic of Korea

The grasses of the *Andropogoneae* tribe—maize, *Sorghum*, sugarcane, and *Miscanthus*—are among the world's most economically important crops, contributing to the food, feed and bioenergy economy. An abundance of genomic resources exist for the two annual crops in this group, maize and *Sorghum* and much is known about their biology. In contrast, genomics studies for the perennial sugarcane and *Miscanthus* have lagged behind, in part because of the size and complexity of their genomes. Using deep sequencing approaches applied to whole genome shotgun sequencing, fosmid inserts, and the transcriptome, we have rapidly gained detailed knowledge of genome organization and gene expression in *Miscanthus*. Using SNP markers identified by RNASeq, we generated a complete genetic map for *Miscanthus sinensis* and discovered that it harbors a recent whole-genome duplication where both genomes retain extensive collinearity with *Sorghum*. A detailed expression atlas of *Miscanthus* (50 billion bases in short reads) reveals a number of biological pathways associated with the unique biology of perennial grasses, in particular the subterranean rhizome, roots, buds and emerging shoots. These data provide a better understanding of the biological activities of the underground stem structure that is the basis for perenniality and the storage or remobilization of nutrient resources in *Miscanthus.* Our efforts also lay the foundation for a *Miscanthus* genome assembly that resolves its distinct sub-genomes and is integrated with a high-density genetic map. These genomics tools will enhance comparative genomics among the highly productive grasses of the *Andropogoneae* tribe and enable future genomics-directed improvement.

Nitrogen assimilation pathways in Arctic microalgae

Ramon Terrado,[1] Adam Monier,[1,2] and Connie Lovejoy* (connie.lovejoy@bio.ulaval.ca)[1,2]

[1]Département de Biologie, Institut de Biologie Intégrative et des Systemes (IBIS) and Québec Océan, Université Laval, Québec, QC, Canada; [2]Takuvik Joint International Laboratory, Université Laval, Canada – Centre National de la Recherche Scientifique, France

Nitrogen, a building block of biological molecules such as nucleotides and amino acids, is a key nutrient for algal growth and development, and is the limiting nutrient for much of the Worlds Ocean. In the ocean, nitrogen is available in numerous dissolved inorganic and dissolved organic forms. In an effort to better understand how different algal species obtain and assimilate nitrogen, we analysed the transcriptomes of five phylogenetically distant microalgae to identify key enzymes for nitrogen transport and assimilation. As much of the Arctic Ocean in particular is depleted in inorganic nitrogen, our DOE-JGI supported project has focused on five isolates from the Arctic Ocean. Cells were grown in standard media supplied with both ammonia and nitrate. Our results show that while all five algae expressed genes for the assimilation of ammonium, not all of them expressed the set of genes required for nitrate assimilation. Also, the expression of genes for the assimilation of organic forms of nitrogen such as urea, nucleosides and small amides was found in some but not all of the microalgae. The transcriptomes of these five Arctic microalgae unveiled their potential for the assimilation of nitrogen from different pools of organic and inorganic nitrogen. We conclude that the algae have evolved various strategies to meet their nitrogen requirements and to assimilate different forms of nitrogen. As a consequence, the relative availability of nitrogen in its varied forms, under changing conditions, is likely to influence the composition of the pelagic microbial community.

Diversity and ecology of Lacustrine haptophyte algae in Lake George, ND, USA: Implications for paleothermometry

Susanna Theroux* (stheroux@lbl.gov),[1,2,3] Yongsong Huang,[2] and Linda Amaral-Zettler[2,3]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Department of Geological Sciences, Brown University, Providence, Rhode Island; [3]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachussets

Lacustrine alkenone records have potential to be valuable sedimentary archives of continental paleotemperature. However, the use of the alkenone-based Uk37 paleotemperature proxy in lake environments is constrained by the genetic diversity of lake-dwelling, alkenone-producing haptophytes. Previous research in Lake George, ND revealed the presence of two alkenone-producing haptophyte species (Hap-A and Hap-B) whose individual contributions to the alkenone sediment record were unknown. To gauge the seasonal abundance of these multiple haptophyte species we used a high-throughput DNA sequencing approach to monitor microbial community composition over the course of the seasonal cycle. Using Ion Torrent sequencing of the18S rRNA gene to determine species identity, we compared water sample microbial communities with water sample alkenone signatures. Additionally, we cultivated Lake George haptophyte isolates in pure and mixed cultures to define their Uk37 temperature calibrations. During the course of the seasonal cycle, total concentrations of alkenones demonstrated a bimodal distribution in the photic zone. The preliminary alkenone peak was characterized by abundant tetraunsaturated (C37:4) alkenones, versus the secondary alkenone peak with abundant triunsaturated (C37:3) alkenones. This variation in water column alkenone signature was reflected in the relative abundance of Hap-A and Hap-B sequence tags. Our culture work determined that these multiple haptophyte isolates required individual Uk37 calibrations that differ from the Lake George *in situ* Uk37 calibration. Lake George sediment alkenone records are therefore composites of multiple, co-occurring haptophyte temperature records. This study is the first next-generation DNA sequencing effort to analyze the microbial community during a haptophyte bloom, and together with culture work, yields a comprehensive understanding of how alkenone signatures in the water column reflect variations in haptophyte species compositions. Our results showcase the genetic predestination of alkenone lipid production and the intricacies of competing temperature records in a lake environment.

Carbon source and light dependent regulation of gene clusters in *Trichoderma reesei (Hypocrea jecorina)*

Doris Tisch, Andre Schuster, and **Monika Schmoll*** (monika.schmoll@ait.ac.at)

AIT Austrian Institute of Technology GmbH, Department Health and Environment, Bioresources, Tulln, Austria

Trichoderma reesei (anamorph of *Hypocrea jecorina*) is one of the most prolific producers of plant cell wall degrading enzymes. Regulation of the genes encoding these enzymes occurs in response to the nutrient sources available in the environment and many of them are responsive to light as well. Cellulose as the natural substrate induces the most complete enzyme set, while induction of cellulases also occurs on sophorose and lactose. In contrast, no cellulases are induced on glycerol and the respective genes are repressed on glucose. We therefore investigated the transcriptome on these five carbon sources in light and darkness and aimed to identify genes specifically expressed under cellulase inducing conditions. These conditions are characterized by a significant enrichment of genes involved in C-compound and carbohydrate degradation and transport among the upregulated gene set. Genes down-regulated under inducing conditions show a significant enrichment in amino acid metabolism, energy metabolism and genes encoding ribosomal proteins and such involved in ribosome biogenesis. We were further interested whether light dependent regulation is clustered in the genome and if the carbon source is relevant for activation of light dependent clusters. We found that light dependent clustering predominantly occurs upon growth on cellulose, with the most significant regulation in a gene cluster comprising env1. This cluster appears on glucose as well, but is not down regulated in mutants of blr1 or blr2. Also cbh2, the arabinofuranosidase gene abf2 and the histone acetyltransferase gene gcn5 are part of light dependent clusters. Hierarchical clustering of gene expression patterns was performed to reveal functional divergence of gene regulation with respect to light response or carbon specific regulation. Glycoside hydrolase genes follow the whole transcriptome pattern with carbon source being superior to light in terms of regulation. However, env1 was found to be crucial for carbon source specific regulation of G-protein coupled receptors, genes involved in secretion, sulphur metabolism and oxidative processes as well as transporters. The g-protein beta subunit gnb1 shows a similar characteristic for GPCRs. We conclude that clustered regulation of light responsive genes preferentially occurs upon growth on the natural carbon source cellulose and that ENV1 and to a lesser extent GNB1 play a role in carbon source dependent regulation of specific gene groups in light.

Evaluation of hypervariable 16S and ITS tag sequencing on Illumina MiSeq.

Julien Tremblay* (jtremblay@lbl.gov),[1] Kanwar Singh,[1] Alison Fern,[1] Edward S. Kirton,[1] Feng Chen,[1] Shaomei He,[1] Tanja Woyke,[1] Janey Lee,[1] Robin A. Ohm,[1] Matthias Hess,[1,2,3] and Susannah G. Tringe[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Systems Biology & Applied Microbial Genomics Laboratory, Washington State University, Richland, Washington; [3]Chemical and Biological Process Development Group, Pacific Northwest National Laboratory, Richland, Washington

In recent years, microbial community surveys extensively relied on 454 pyrosequencing technology (pyrotags). The Illumina sequencing platform HiSeq2000 has now largely surpassed 454 in terms of read quantity and quality with typical yields of up to 600 Gb of paired-end 150 bases reads in one 18 day run. Illumina recently introduced the new mid-range MiSeq sequencing platform which gives an output of more than 1 Gb of paired-end 250 base reads in a single day run. With its moderately-high throughput and support for high multiplexing (barcoding), this platform represents a suitable alternative for projects needing more conservative throughput, for instance population profiling based on prokaryotic 16S rRNA genes or eukaryotic Internal Transcribed Spacer (ITS) regions. A workflow was therefore developed to take advantage of the Illumina MiSeq platform as a suitable tool to accurately characterize microbial communities. We surveyed microbial populations coming from various environments by targeting the 16S rRNA hypervariable regions V4, V7-V8 and V6-V8 which generated amplicons size of about 290, 307 and 460 bp respectively. These amplicons were sequenced with a MiSeq instrument from both 5' and 3' ends with a 2x250 bases sequencing configuration followed by *in silico* assembly using their shared overlapping part when applicable. Corresponding V6-V8 pyrotags data were also generated to assess validity of itag data. We also explored amplicons sequencing of the ITS2 region and examined how this marker gene can be used for fungal profiling. Although it generates shorter reads than the 454 platform, MiSeq combines well balanced throughput with a remarkably low error rate. Our results suggest that the itags community surveys on MiSeq successfully recapture known biological results and should provide a useful tool for both prokaryotic and eukaryotic community characterization.

Full length cDNA sequencing on the PacBio® *RS*

Elizabeth Tseng* (etseng@pacificbiosciences.com)[1] and Jason G. Underwood[1]

[1]Pacific Biosciences, Menlo Park, California

Transcriptome sequencing using short read technologies (RNA-seq) provides valuable information on transcript abundance and rare transcripts. While short reads can be used to infer alternative splicing and variable transcription start sites, the use of short reads for these research questions creates computational problems due to uneven read coverage, complex splicing and potential sequencing bias. Here, we demonstrate the long readlength capabilities of the PacBio® *RS* to sequence full-length cDNA molecules derived from human polyA RNA. Our library preparation method generates sequencing libraries highly enriched in full-length cDNA molecules. Because the PacBio® *RS* uses a circular sequencing structure, reads are putatively full-length if either both ends of the SMRT adapter or both the 5' and 3' cDNA library adaptor primers are seen. By mapping putative full-length reads against the Gencode database, we show that we recovered many full-length transcripts spanning a range of 500 – 6,000 bp in length. In addition, we identified potential alternative isoforms of known genes as well as novel genes. We also describe two published error correction methods, PacBioToCA and LSC, for improving PacBio read accuracy using Illumina short reads. We report our findings on detecting novel splicing events and full-length transcript characterization in a human sample, showing that PacBio® *RS* sequencing technology can assist researchers in better characterizing the transcriptome in its native, full-length form and help unlock combinatorial RNA processing regulation not observed in previous RNA-seq experiments.

Structural variation detection and *de novo* assembly in complex genomes using extremely long single-molecule imaging

H. VanSteenhouse* (hvansteenhouse@bionanogenomics.com), A. Hastie, E. Lam, H. Dai, M. Requa, M. Austin, F. Trintchouk, M. Saghbini, and H. Cao

BioNano Genomics, San Diego, California

De novo genome assemblies using only short read data are generally incomplete and highly fragmented due to the intractable complexity found in most genomes. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses. We present a single molecule genome analysis system (Irys) based on NanoChannel Array technology that linearizes extremely long DNA molecules for observation. This high-throughput platform automates the imaging of single molecules of genomic DNA hundreds of kilobases in size to measure sufficient sequence uniqueness for unambiguous assembly of complex genomes. High-resolution genome maps assembled *de novo* from the extremely long single molecules retain the original context and architecture of the genome, making them extremely useful for structural variation and assembly applications. Genome map-based scaffolding in shotgun sequencing experiments performed in parallel with second or third generation sequence production offers an integrated pipeline for whole genome *de novo* assembly solving many of the ambiguities inherent when using sequencing alone. Additionally, genome maps serve as a much-needed orthogonal validation method to NGS assemblies. As a result, genome maps improve contiguity and accuracy of whole genome assemblies, permitting a more comprehensive analysis of functional genome biology and structural variation. In addition to providing an introduction to this newly available technology, we will demonstrate a number of examples of its utility in a variety of organisms, including an arthropod, fungus, and crop plant.

The DOE Systems Biology Knowledgebase: Plants Science Domain

Doreen Ware (ware@cshl.edu),[1,2] David Weston,[3] Sergei Maslov,[4] Shinjae Yoo,[4] Dantong Yu,[4] Michael Schatz,[1] James Gurtowski,[1] Matt Titmus,[1] Jer-ming Chia,[1] Sunita Kumari,[1] Andrew Olson,[1] Shiran Pasternak,[1] Jim Thomason,[1] Ken Youens-Clark,[1] Mark Gerstein,[5] Gang Fang,[5] Daifeng Wang,[5] Pam Ronald,[6] TaeYun Oh,[6] Chris Henry,[7] Sam Seaver,[7] **Priya Ranjan***,[3] Mustafa Syed,[3] Miriam Land,[3] and Adam Arkin[8]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [2]U.S. Department of Agriculture – Agriculture Research Service; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Brookhaven National Laboratory, Upton, New York; [5]Yale University, New Haven, Connecticut; [6]University of California at Davis, Davis, California; [7]Argonne National Laboratory, Argonne, Illinois; [8]Lawrence Berkeley National Laboratory, Berkeley, California

The Department of Energy Systems Biology Knowledgebase (KBase) is a software and data environment designed to enable researchers to collaboratively generate, test and share new hypotheses about gene and protein functions; perform large-scale analyses on our scalable computing infrastructure; and model interactions in microbes, plants, and their communities. It permits secure sharing of data, tools, and

scientific conclusions in a unified and extensible framework that does not require users to learn separate systems. The major goal for the KBase plants area is to model genotype-to-phenotype relationships through analysis and integration of diverse 'omics data sets. These include genomic, transcriptomic, methylomic sequencing, metabolite and phenotype measurements, and the reconstruction of metabolic and functional networks based on expression profiles, protein-DNA, and protein-protein interactions. The KBase Plants team will provide a suite of services to achieve this goal in the long term. Currently released preliminary services include a jnomics-based genotyping workflow service, a data exploration and visualization service for results of genome wide association studies, and co-expression network analysis tools and metabolic reconstruction tools.

Intra-genus diversity of *Bradyrhizobium spp.* isolated from bulk forest soil — an examination of free-living versus nodulating lifestyles

Roland Wilhelm* (rwilhelm@mail.ubc.ca)**,** David VanInsberghe, Steven Hallam, and William W. Mohn

Department of Microbiology and Immunology, Life Sciences Institute, University of British Columbia, Vancouver, Canada

The original description of the genus *Bradyrhizobium* as nodulating, nitrogen fixing root symbionts is complicated by evidence of a diverse group of free-living species (incapable of nodulation) and non-nodulating plant symbionts whose ecological role is poorly understood . Comprehensive rRNA gene pyrotag analysis of bacterial communities in bulk soil of six softwood forest sites in British Columbia demonstrated that a clade of closely related *Bradyrhizobium spp.* were the most abundant taxa, constituting 15% of all classifiable reads (981,472). Their prevalence in bulk-soil and lack of an apparent host suggest that these members of *Bradyrhizobium* are unlikely involved in classic nodulation, because such strains typically show host-specificity. The functions of this clade in forest soils are unknown but likely important. With nine strains from within this clade in culture, we have begun characterizing their metabolic and possible ecological functions by evaluating carbon, nitrogen, sulphur and phosphorous metabolism and by comparative analysis of their genomes. We will identify characteristics of free-living versus symbiotic lifestyles in these bulk-soil *Bradyrhizobium* strains via comparison to known nodulating strains, such as *Bradyrhizobium japonicum*. Our first two draft genomes both lacks *nif* genes, involved in nitrogen fixation, as well as the classic set of *nod* genes, involved in root nodulation, suggestive of a free-living lifestyle. The presence of non-canonical *nod* genes and a variety of enzymes involved in synthesis or modification of phytohormones suggest some level of plant-interaction. This research will improve understanding of the diversity within the family *Bradyrhizobaceae* and determine the plasticity and diversity of symbiotic and free-living lifestyles within this genus. It will also help characterize the function of this abundant clade in softwood forest soil communities, which will improve our ecological understanding of this important ecosystem.

Phylogenetic tree based taxonomic classification

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

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

Our current understanding of the taxonomic and phylogenetic diversity of cellular organisms, especially the bacteria and archaea, is mostly based upon studies of sequences of the ribosomal RNA gene sequences, especially those for the small-subunit rRNA (ss-rRNA). The current taxonomic classification of bacteria and archaea is also heavily based on ssu-rRNA. Despite the historical and current power of ssu-rRNA analysis, it does have some drawbacks including copy number variation among organism and complications introduced by horizontal gene transfer, convergent evolution, or evolution rate variations. Fortunately, genome sequencing and metagenomic sequencing are providing a wealth of information about other genes in the genomes of various bacteria and archaea. By analyzing complete genome sequences in the IMG database, we have identified 40 protein-coding genes with strong potential as broad phylogenetic markers across bacteria and archaea (e.g., they are highly universal, have low variation in copy number, and have relatively congruent phylgoenetic trees). We report here the development and use of methods to make use of these 40 phylogenetic marker genes for operational taxonomic unit assignment and taxonomic classification of bacteria and archaea. Our method allows one to place an organism into a specific taxonomic group at various taxonomic levels while accounting for differences in rates of evolution between taxa and between genes. We compare the OTUs and taxonomic classifications for these protein coding marker genes with OTUs and classifications based on phylogenetic trees of ss-rRNA and those from sequence clustering (non phylogenetic) methods. Our analysis demonstrates that, at the species level, phylogenetic tree-based methods examining these 40 protein coding genes identify OTUs that are comparable to ss-rRNA sequence similarity based OTUs. Our phylogenetic tree based taxonomic classifications of IMG genomes at the genus, order, family, class, phylum levels will be discussed.

Transformation of *Brachypodium distachyon* to analyze CNS enhancer activity and genome fractionation in maize

Hugh Young* (hugh.young@ars.usda.gov),**[1,2]** James Schnable,[3] Michael Freeling,[2] and John Vogel[1]

[1]U.S. Department of Agriculture-Agricultural Research Service-Western Regional Research Center, Albany, California; [2]University of California at Berkeley, Berkeley, California; [3]Donald Danforth Center, St. Louis, Missouri

Fractionation is the deletion of one or the other but not both of a pair of homeologous genes following an ancient whole genome duplication. In maize, these deletion events result from intra-chromosomal recombination. Analyses of expression data (RNAseq) between homeologous gene pairs in maize revealed that genes from maize1 (the least fractionated subgenome) tended to be expressed more than duplicates of the same genes from maize2. Significant differences between maize1 to maize2 expression have been found to occur in specific tissues, such as maize pollen. To determine the polymorphisms responsible for these expression differences, we have compared the noncoding sequences surrounding pairs of maize genes which showed different patterns of expression. Syntenic orthologs from other grass species (sorghum, foxtail millet, rice, Brachypodium) are included in the analysis to identify conserved non-coding sequences (CNSs) which are functionally constrained and may act in the regulation of gene expression. In order to test the enhancer activity of these CNSs, several have been cloned into the binary vector pGPro8 (GenBank JN593327), which is modified to include a minimal 35S promoter (35SMin) ahead of a GUS reporter gene. Here we describe the stable transformation of *Brachypodium distachyon* (Bd21-3) plants to test the activity of pollen-specific CNS enhancers identified through analysis of fractionation mutagenesis in maize. Stable T_0 plants carrying the pGPro8-35SMin vector with potential CNS enhancers demonstrate pollen and anther-specific GUS expression.

Lignocellulose-derived inhibitors from ammonia pre-treated biomass activate specific regulatory circuits and inhibit bacterial conversion of xylose to ethanol

Yaoping Zhang,[1] David Keating,[1] Irene Ong,[1] Sean McIlwain* (smcilwain@glbrc.wisc.edu),[1] Jeff Grass,[1] Donna Bates,[1] Alan Higbee,[1] Josh Coon,[1] Tricia Kiley,[1] Yury Bukhman,[1] Mingjie Jin,[2] Ven Balan,[2] Bruce Dale,[2] Mary Lipton,[3] Josh Aldrich,[3] and Bob Landick[1]

[1]Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wisconsin; [2]Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan; [3]Pacific Northwest National Laboratory, Richland, Washington

Lignocellulose-derived inhibitors of microbial physiology and metabolism generate major barriers to efficient conversion of sugars to biofuels. The complex chemical composition of biomass hydrolysates, however, makes it difficult to identify cellular targets of these inhibitors and the mechanism of their effects on microbial sugar conversion. We have developed an approach to study these questions in hydrolysates of alkaline-pretreated biomass, specifically AFEX-pretreated corn stover hydrolysate (ACSH), by devising a chemically defined synthetic hydrolysate that largely replicates the properties of ACSH including a cocktail of LTs identified in ACSH. We used a multiomic dissection of the effects of synH and LTs on an engineered *Escherichia coli* ethanologen to identify key stress responses in ACSH caused by LTs. We executed this study in two phases. An initial phase defined version 1 of synH, whose comparison to ACSH allowed identification of osmolytes and LTs as the key missing components of a complete synH (Schwalbach *et al*., 2012 *AEM* 78:3442-57). Based on knowledge obtained from that study, SynH version 3 (SynHv3) was developed that more faithfully represents the composition of ACSH and included chemically synthesized LTs specific for hydrolysates derived from AFEX-pretreated biomass. SynHv3 allowed us to test the effects of LTs on conversion, independent of other stress-inducing components. Multiomic analysis indicated that SynHv3 + LTs largely recapitulates the growth and physiology of *E. coli* grown in ACSH. Using it, we identified 4 major stress regulons of high relevance to ACSH and LT-inhibition of xylose conversion, the AaeR, MarR/MarA, YqhC, and FrmR regulons. We also found that nearly all the cellular functions upregulated as a stress response to LTs occurred at the transcriptional level, rather than by translational regulation. These stress responses were associated with a depletion of NADPH/NADH, inefficient ethanol synthesis, and pyruvate pooling. Two aldehyde stress responses (YqhC and FrmR regulons) were ameliorated when cells entered a stationary phase during which most xylose conversion occurs. However, the AaeR and MarR/MarA regulons, which govern efflux pumps in addition to detoxification genes, remained active in stationary phase suggesting that continued intracellular presence of toxic small molecules/metabolites may be responsible for inhibition of xylose conversion.

PacBio only assembly with low genomic DNA input

Zhiying Zhao* (zyzhao@lbl.gov),[1] Yu-Chih Tsai,[2] Alicia Clum,[1] Katherine Munson,[1] Chris Daum,[1] Stephen W. Turner,[2] Jonas Korlach,[2] Len A. Pennacchio,[1] and Feng Chen[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Pacific Biosciences, Menlo Park, California

At JGI, we use PacBio single molecule DNA sequencing as a quick-turnaround and cost-effective solution for drafting and finishing microbial genomes that remains a largely unexplored area of life. Construction of PacBio library by traditional protocol still requires micrograms' genomic DNA. In many cases, getting high quantity of genomic DNA remains as a major challenge. Recently, PacBio developed a more efficient library construction method using terminal deoxynucleotidyl transferase (TdT), which makes it possible to obtain sufficient sequencing data for assembly from significantly smaller amount of genomic DNA. We have tested and validated this newly developed method. Sequencing results will be presented. We also tested PacBio data only assembly approach in combination with this library construction protocol. The analysis results will be presented as well.

Attendees

Current as of March 14, 2013

Timothy Alba
University of Nevada, Las Vegas
albat@unlv.nevada.edu

Eric Allen
University of California, San Diego
eallen@ucsd.edu

Dilara Ally
SG Biofuels
dally@sgbiofuels.com

Rick Amasino
University of Wisconsin
amasino@biochem.wisc.edu

George Anasontzis
Chalmers University of Technology
george.anasontzis@chalmers.se

William Andreopoulos
DOE Joint Genome Institute
wandreopoulos@lbl.gov

Emma Aronson
University of California, Irvine
emma.aronson@gmail.com

Viridiana Avila
University of California, Merced
efectodoppler.viridiana@gmail.com

Charles Bachy
Monterey Bay Aquarium Res Institute
cbachy@mbari.org

Massie Ballon
DOE Joint Genome Institute
mlballon@lbl.gov

Richard Baran
Lawrence Berkeley National Lab
RBaran@lbl.gov

Danny Barash
Ben-Gurion University
dbarash@cs.bgu.ac.il

Kerrie Barry
DOE Joint Genome Institute
kwbarry@lbl.gov

Sajeev Batra
DOE Joint Genome Institute
sbatra@lbl.gov

Diane Bauer
DOE Joint Genome Institute
dmbauer@lbl.gov

Chris Beecroft
DOE Joint Genome Institute
cjbeecroft@lbl.gov

Gregory Bell
Lawrence Berkeley National Lab
grbell@lbl.gov

Renaud Berlemont
University of California, Irvine
rberlemo@ucl.edu

Eldredge Bermingham
Smithsonian Tropical Res Institute
bermingham@si.edu

Paul Blainey
The Broad Institute
pblainey@broadinstitute.org

Jeffrey Blanchard
University of Massachusetts
jeffb@bio.umass.edu

Matthew Blow
DOE Joint Genome Institute
mjblow@lbl.gov

Barry Bochner
Biolog, Inc.
bbochner@biolog.com

Benjamin Bower
Dupont Industrial Biosciences
ben.bower@dupont.com

Chris Bowler
Ecole Normale Superieure
cbowler@biologie.ens.fr

Alexander Boyd
DOE Joint Genome Institute
aeboyd@lbl.gov

Lambert Brau
Deakin University
lambert.brau@deakin.edu.au

Susan Brawley
Carnegie Institution
sbrawley@stanford.edu

Natalie Breakfield
Univ. of North Carolina- Chapel Hill
nbreakfield@gmail.com

Eoin Brodie
Lawrence Berkeley National Lab
elbrodie@lbl.gov

Natasha Brown
DOE Joint Genome Institute
nzvenigorodsky@lbl.gov

Yury Bukhman
Great Lakes Bioenergy Res Center
ybukhman@glbrc.wisc.edu

Frank Burns
DuPont
frank.r.burns@usa.dupont.com

Brian Bushnell
DOE Joint Genome Institute
bbushnell@lbl.gov

Gregory Butler
Concordia University
gregb@cs.concordia.ca

Hong Cai
New York University Abu Dhabi
hc55@nyu.edu

Douglas Cameron
First Green Partners
dcc@firstgreenpartners.com

Shane Canon
LBNL/NERSC
scanon@lbl.gov

Erick Cardenas
Poire University of British
Columbiacarden24@mail.ubc.ca

Joe Carlson
US DOE Joint Genome Institute
JWCarlson@lbl.gov

David Cavalier
Great Lakes Bioenergy Res Center
cavalie8@msu.edu

Leong Keat Chan
DOE Joint Genome Institute
leongchan@lbl.gov

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

Cindy Chen
DOE Joint Genome Institute
cindychen@lbl.gov

Feng Chen
DOE Joint Genome Institute
fchen@lbl.gov

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

Jason Chin
Pacific Biosciences
jchin@pacificbiosciences.com

Jennifer Chiniquy
DOE Joint Genome Institute
JLChiniquy@lbl.gov

Penny Chisholm
MIT
chisholm@mit.edu

Dylan Chivian
Lawrence Berkeley National Lab
DCChivian@lbl.gov

In-Geol Choi
Korea University
igchoi@korea.ac.kr

Mansi Chovatia
DOE Joint Genome Institute
mrchovatia@lbl.gov

George Chuck
University of California, Berkeley
georgechuck@berkeley.edu

Doina Ciobanu
DOE Joint Genome Institute
dgciobanu@lbl.gov

Milica Ciric
AgResearch Ltd.
Milica.Ciric@agresearch.co.nz

Scott Clingenpeel
DOE Joint Genome Institute
srclingenpeel@lbl.gov

Alicia Clum
DOE Joint Genome Institute
aclum@lbl.gov

Devin Coleman-Derr
DOE Joint Genome Institute
dacoleman-derr@lbl.gov

Roy Eric Collins
University of Alaska Fairbanks
rec3141@gmail.com

Alex Copeland
DOE Joint Genome Institute
accopeland@lbl.gov

Ciara Curtin
GenomeWeb
ccurtin@genomeweb.com

Chris Daum
DOE Joint Genome Institute
cgdaum@lbl.gov

Anne Dekas
Lawrence Livermore National Lab
dekas1@llnl.gov

Joseph DeRisi
UCSF / HHMI
joe@derisilab.ucsf.edu

Andreas Desiniotis
University of Athens
a_desiniotis@hotmail.com

Sam Deutsch
DOE Joint Genome Institute
SDeutsch@lbl.gov

Anthony Devine
RTI International
adevine@rti.org

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

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

Kecia Duffy
DOE Joint Genome Institute
kmduffy@lbl.gov

Rob Egan
DOE Joint Genome Institute
RSEgan@lbl.gov

John Eid
Pacific Biosciences
jeid@pacificbiosciences.com

Hamza El Dorry
The American University in Cairo
dorry@aucegypt.edu

Marsha Fenner
DOE Joint Genome Institute
MWFenner@lbl.gov

James Fisher
Life Technologies
james.fisher@lifetech.com

Brian Foster
DOE Joint Genome Institute
bfoster@lbl.gov

Mohamed Ghazy
The American University in Cairo
mghazy@aucegypt.edu

M Thomas P Gilbert
Natural History Museum of Denmark
mtpgilbert@gmail.com

David Gilbert
DOE Joint Genome Institute
degilbert@lbl.gov

Celeste Glazer
Labcyte
cglazer@labcyte.com

Lynne Goodwin
Los Alamos National Laboratory
lynneg@lanl.gov

Sean Gordon
USDA
seangordon07@gmail.com

Alex Greenspan
University of California, Davis
greenspan@ucdavis.edu

Igor Grigoriev
DOE Joint Genome Institute
ivgrigoriev@lbl.gov

Andrey Grigoriev
Rutgers University
agrigoriev@camden.rutgers.edu

Stephen Gross
DOE Joint Genome Institute
smgross@lbl.gov

Jenny Gu
Pacific Biosciences
jgu@pacificbiosciences.com

Adam Guss
Oak Ridge National Laboratory
gussam@ornl.gov

Masood Hadi
NASA
Masood.hadi@nasa.gov

James Han
DOE Joint Genome Institute
jkhan@lbl.gov

Sajeet Haridas
DOE Joint Genome Institute
sharidas@lbl.gov

Miranda Harmon-Smith
DOE Joint Genome Institute
MLHarmon-Smith@lbl.gov

Alyse Hawley
University of British Columbia
alysekh@mail.ubc.ca

Erik Hawley
Washington State University
ehawley.amge@gmail.com

Matthew Haynes
DOE Joint Genome Institute
mrhaynes@lbl.gov

Sam Hazen
University of Massachusetts
hazen@bio.umass.edu

Guillermina Hernandez-Raquet
Institut Nation de la Recherche
hernandg@insa-toulouse.fr

Sur Herrera Paredes
UNC-CH
sur00mx@gmail.com

Matthias Hess
Washington State University
matthias.hess@tricity.wsu.edu

Robert Hettich
Oak Ridge National Laboratory
hettichrl@ornl.gov

Luke Hickey
Pacific Biosciences
LHickey@pacificbiosciences.com

Nathan Hillson
Joint BioEnergy Institute
njhillson@lbl.gov

Jennifer Hiras
Joint BioEnergy Institute
jhiras@lbl.gov

Cindi Hoover
DOE Joint Genome Institute
cahoover@lbl.gov

Susan Hua
DOE Joint Genome Institute
shua@lbl.gov

Amy Huang
DOE Joint Genome Institute
amyhuang@lbl.gov

Laura Hug
University of California, Berkeley
laura.hug@berkeley.edu

Ashish Kumar Jaiswal
New York University Abu Dhabi
akj4@nyu.edu

Janet Jansson
Lawrence Berkeley National Lab
jrjansson@lbl.gov

Nicole Johnson
DOE Joint Genome Institute
NicoleVJohnson@lbl.gov

Richard Jorgensen
Frontiers in Plant Science
rajorgensen@mac.com

Chijioke Joshua
LBNL/JBEI
cjjoshua@lbl.gov

Dongwan Kang
Lawrence Berkeley National Lab
ddkang@lbl.gov

Ulas Karaoz
Lawrence Berkeley National Lab
ukaraoz@lbl.gov

Achim Karger
Life Technologies
achim.karger@lifetech.com

Peter Karp
SRI International
pkarp@ai.sri.com

Eric Karsenti
EMBL
karsenti@embl.de

Lisa Kegg
DOE Joint Genome Institute
lrkegg@lbl.gov

Katharina Keiblinger
Univ. of Natural Resources & Life Sciences Vienna
katharina.keiblinger@boku.ac.at

Megan Kennedy
DOE Joint Genome Institute
mckennedy@lbl.gov

Richard Kerrigan
Sylvan Biosciences
rwk@sylvaninc.com

Shintaro Kikuchi
Muroran Institute of Technolog
shintaro@mmm.muroran-it.ac.jp

Jeff Kimbrel
JBEI/LBL
jakimbrel@lbl.gov

Edward Kirton
DOE Joint Genome Institute
eskirton@lbl.gov

Kirill Krivushin
University of Toronto
kirill.krivushin@utoronto.ca

Alan Kuo
DOE Joint Genome Institute
akuo@lbl.gov

Rita Kuo
Lawrence Berkeley National Lab
rckuo@lbl.gov

Kurt LaButti
DOE Joint Genome Institute
klabutti@lbl.gov

David Lai
DOE Joint Genome Institute
dlai@lbl.gov

Kathleen Lail
DOE Joint Genome Institute
klail@lbl.gov

Sadhana Lal
University of Manitoba
sadhana.lal@gmail.com

Peter Larsen
Argonne National Laboratory
plarsen@anl.gov

Jane Lau
JBEI-LBNL
jlau@lbl.gov

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

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

Charles Lee
University of Waikato
cklee@waikato.ac.nz

Jackson Lee
NASA Ames Research Center
jackson.z.lee@nasa.gov

Denis Legault
Concordia University
denis.legault@concordia.ca

Emily Leproust
Agilent Technologies
emily_leproust@agilent.com

Hilary Leung
University of British Columbia
hilaryl@mail.ubc.ca

Konstantinos Liolios
DOE Joint Genome Institute
kliolios@lbl.gov

Anna Lipzen
DOE Joint Genome Institute
alipzen@lbl.gov

Elizabeth Lobos
DOE Joint Genome Institute
ealobos@lbl.gov

Dominique Loque
JBEI - LBNL
dloque@lbl.gov

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

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

Derek Lundberg
UNC Chapel Hill
derek.lundberg@gmail.com

Stephanie Malfatti
DOE Joint Genome Institute
samalfatti@lbl.gov

Rex Malmstrom
DOE Joint Genome Institute
rrmalmstrom@lbl.gov

Joel Martin
DOE Joint Genome Institute
j_martin@lbl.gov

Eric Mathur
SG Biofuels
emathur@sgbiofuels.com

Sean McIlwain
Great Lakes Bioenergy Res Center
smcilwain@glbrc.wisc.edu

Xiandong Meng
DOE Joint Genome Institute
xiandongmeng@lbl.gov

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

Sirma Mihaltcheva
DOE Joint Genome Institute
smihaltcheva@lbl.gov

Therese Mitros
DOE Joint Genome Institute
tkmitros@lbl.gov

Supratim Mukherjee
DOE Joint Genome Institute
supratimmukherjee@lbl.gov

Aindrila Mukhopadhyay
Lawrence Berkeley National Lab
amukhopadhyay@lbl.gov

Christopher Mungall
Lawrence Berkeley National Lab
cjmungall@lbl.gov

Monica Munoz-Torres
Lawrence Berkeley National Lab
mcmunozt@lbl.gov

Katherine Munson
DOE Joint Genome Institute
kmmunson@lbl.gov

Senthil Murugapiran
University of Nevada, Las Vegas
senthil.murugapiran@unlv.edu

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

Niranjan Nagarajan
Genome Institute of Singapore
nagarajann@gis.a-star.edu.sg

li Navid
Lawrence Livermore National Lab
navid1@llnl.gov

David Nelson
New York University
davidroynelson@gmail.com

Jack Newman
Amyris
Newman@amyris.com

Chewyee Ngan
DOE Joint Genome Institute
cyngan@lbl.gov

Matt Nolan
DOE Joint Genome
Institutempnolan@lbl.gov

Angela Norbeck
EMSL, Pacific Northwest National Lab
angela.norbeck@pnnl.gov

Trent Northen
Lawrence Berkeley National Lab
trnorthen@lbl.gov

Bahador Nosrat
DOE Joint Genome Institute
bnosrat@lbl.gov

Aki Ohdera
University of California, Merced
aohdera@ucmerced.edu

Robin Ohm
DOE Joint Genome Institute
raohm@lbl.gov

Jose Olivares
Los Alamos National Laboratory
olivares@lanl.gov

Christian Olsen
Biomatters
Christian@biomatters.com

Bobby Otillar
DOE Joint Genome Institute
rpotillar@lbl.gov

Tatiana Paley
DOE Joint Genome Institute
tlpaley@lbl.gov

Terry Payette
Luca Technologies
Terry.Payette@lucatechnologies.com

Paul Peluso
Pacific Biosciences
ppeluso@pacificbiosciences.com

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

Jennifer Pett-Ridge
Lawrence Livermore National Lab
pettridge2@llnl.gov

Hailan Piao
Washington State University
hitzroth@tricity.wsu.edu

Faradia Pierre
DOE Joint Genome Institute
fpierre@lbl.gov

Emmanuel Prestat
Lawrence Berkeley National Lab
EPrestat@lbl.gov

Jose Pruneda-Paz
University of California, San Diego
jprunedapaz@ucsd.edu

Priya Ranjan
University of Tennessee, Knoxville
pranjan@utk.edu

David Rank
Pacific Biosciences
drank@pacificbiosciences.com

Igor Ratnere
DOE Joint Genome Institute
iratnere@lbl.gov

Tatiparthi Reddy
DOE Joint Genome Institute
tbreddy@lbl.gov

Wayne Reeve
Murdoch University
w.reeve@murdoch.edu.au

Amanda Reider Apel
JBEI
arreiderapel@lbl.gov

Sarah Richardson
DOE Joint Genome Institute
smrichardson@lbl.gov

Barbara Ruef
Maverix Biomics
bruef@maverixbio.com

Asaf Salamov
DOE Joint Genome Institute
aasalamov@lbl.gov

Laura Sandor
DOE Joint Genome Institute
LCSandor@lbl.gov

Rupinder Sayal
DOE Joint Genome Institute
rsayal@lbl.gov

Wendy Schackwitz
DOE Joint Genome Institute
wsschackwitz@lbl.gov

Christopher Schadt
Oak Ridge National Laboratory
schadtcw@ornl.gov

Alicia Scheffer
Agilent Technologies
alicia_scheffer@agilent.com

Bob Schmidt
SG Biofuels
bschmidt@sgbiofuels.com

Monika Schmoll
Austrian Institute of Technology
monika.schmoll@ait.ac.at

Jeremy Schmutz
HudsonAlpha Institute/JGI
jschmutz@hudsonalpha.org

Duncan Scott
DOE Joint Genome Institute
dnscott@lbl.gov

Nicole Shapiro
DOE Joint Genome Institute
nrshapiro@lbl.gov

Aditi Sharma
DOE Joint Genome Institute
aditisharma@lbl.gov

Shengqiang Shu
DOE Joint Genome Institute
sqshu@lbl.gov

Steven Singer
Lawrence Berkeley National Lab
SWSinger@lbl.gov

Kanwar Singh
DOE Joint Genome Institute
ksingh@lbl.gov

Jeffrey Skerker
University of California, Berkeley
skerker1@gmail.com

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

Tatyana Smirnova
DOE Joint Genome Institute
tsmirnova@lbl.gov

Kristi Spittle
Pacific Biosciences
kspittle@pacificbiosciences.com

Dimitrios Stamatis
DOE Joint Genome Institute
dastamatis@lbl.gov

Ben Stenuit
University of California, Berkeley
bstenuit@berkeley.edu

Solomon Stonebloom
JBEI
sstonebloom@lbl.gov

Rhona Stuart
Lawrence Livermore National Lab
stuart25@llnl.gov

Sebastian Sudek
Monterey Bay Aquarium Res Institute
ssudek@mbari.org

Sirisha Sunkara
DOE Joint Genome Institute
ssunkara@lbl.gov

Kankshita Swaminathan
University of Illinois
kanksw@gmail.com

Ernest Szeto
Lawrence Berkeley National Lab
eszeto@lbl.gov

Shuiquan Tang
University of Toronto
Shuiquan.tang@mail.utoronto.ca

Tootie Tatum
DOE Joint Genome Institute
oltatum@lbl.gov

Kristin Tennessen
DOE Joint Genome Institute
ktennessen@lbl.gov

Susanna Theroux
DOE Joint Genome Institute
stheroux@lbl.gov

Hope Tice
DOE Joint Genome Institute
HNTice@lbl.gov

Julien Tremblay
DOE Joint Genome Institute
jtremblay@lbl.gov

Susannah Tringe
DOE Joint Genome Institute
sgtringe@lbl.gov

Stephan Trong
DOE Joint Genome Institute
strong@lbl.gov

Edward Turano
Lawrence Berkeley National Lab
EJTurano@lbl.gov

Ray Turner
DOE Joint Genome Institute
crturner@lbl.gov

Gerald Tuskan
Oak Ridge National Laboratory
tuskanga@ornl.gov

Jeltje Van Baren
Monterey Bay Aquarium Res Institute
jeltje@mbari.org

Daniel Van der Lelie
RTI International
vdlelied@rti.org

Linda Van Diepen
University of New Hampshire
lvandiep@gmail.com

Harper VanSteenhouse
BioNano Genomics
hvansteenhouse@bionanogenomics.com

Axel Visel
JGI/LBNL
avisel@lbl.gov

Chris Voigt
MIT
cavoigt@gmail.com

Carey Waage
DOE Joint Genome Institute
cvwaage@lbl.gov

Maggie Wagner
Duke University
mrw28@duke.edu

Joe Walp
DOE Joint Genome Institute
jcwalp@lbl.gov

Zhong Wang
DOE Joint Genome Institute
zhongwang@lbl.gov

Jordan Waters
DOE Joint Genome Institute
jwaters@lbl.gov

Roland Wilhelm
University of British Columbia
rwilhelm@mail.ubc.ca

Andreas Wilke
Argonne National Laboratory
wilke@mcs.anl.gov

Patrick Wincker
CEA - Genoscope
pwincker@genoscope.cns.fr

Chee Hong Wong
Joint Genome Institute
chwong@lbl.gov

Alexandra Worden
Monterey Bay Aquarium Res
Instituteazworden@mbari.org

Tanja Woyke
DOE Joint Genome Institute
twoyke@lbl.gov

Dongying Wu
University of California, Davis
dygwu@ucdavis.edu

Yu-Wei Wu
Lawrence Berkeley National Lab
ywwei@lbl.gov

Hugh Young
USDA-ARS-WRRC
hugh.young@ars.usda.gov

Scott Yourstone
University of North Carolina
scott.yourstone81@gmail.com

Alex Hon-Tsen Yu
National Taiwan University
ayu@ntu.edu.tw

Matthew Zane
DOE Joint Genome Institute
mczane@lbl.gov

Yubo Zhang
DOE Joint Genome Institute
yubozhang@lbl.gov

Jean Zhao
DOE Joint Genome Institute
zyzhao@lbl.gov

Gerald Zon
Trilink Biotechnologies
gzon@trilinkbiotech.com

Author Index

JGI
DOE JOINT GENOME INSTITUTE

U.S. DEPARTMENT OF
ENERGY
Office of Science

CSO 25781